GLOBAL DYSREGULATION OF GENE EXPRESSION AND TUMORIGENESIS: DATA SCIENCE FOR CANCER

A Dissertation
Presented to
The Academic Faculty

by

Evan Clayton

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics in the
School of Biological Sciences

Georgia Institute of Technology
December 2019

COPYRIGHT © 2019 BY EVAN CLAYTON

GLOBAL DYSREGULATION OF GENE EXPRESSION AND TUMORIGENESIS: DATA SCIENCE FOR CANCER

Approved by:

Dr. John McDonald, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Jung Choi
School of Biological Sciences
Georgia Institute of Technology

Dr. King Jordan, Co-advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Peng Qiu
Department of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Soojin Yi
School of Biological Sciences
Georgia Institute of Technology

Date Approved: August 22, 2019

To my family and friends

ACKNOWLEDGEMENTS

I am truly grateful to my advisor Dr. John McDonald for his guidance throughout my time working with him as a PhD student. I will always be appreciative of his relentless support and belief in me as a research scientist. His passion for science is tangible and has left a lasting impression on me. Under his tutelage, I have grown greatly as a scientist and independent critical thinker.

I am also very thankful for my co-advisor Dr. King Jordan and his patience for teaching many fundamental scientific concepts. His ability to breakdown complex ideas and explain them in a manner which is comprehendible has aided my scientific knowledge and presentation skills. With his mentorship, I have been exposed to cutting-edge bioinformatics and expanded my breadth of science.

Apart from my direct advisors, I have had the pleasure of working with the inspiring members of my research committee; Dr. Soojin Yi, Dr. Jung Choi, and Dr. Peng Qiu. Their constant support and feedback have been insightful and critical for my success and development.

I am very appreciative for the close friends and colleagues I have made during the course of my PhD. Dr. Lavanya Rishishwar and Dr. Lu Wang were incredible mentors in my early years. Toyya Pujol, Shareef Khalid, Shashwat Deepali and Dongjo Ban have been close friends who significantly impacted my scientific knowledge and research. Their encouragement throughout my PhD has been invaluable.

Lastly, none of this would be possible without my parents Jeffrey Clayton and Virginia Clayton. Their love and support have put me in a position to succeed and allowed me to be where I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

1KGP	1000 Genomes Project
5-FU	Fluorouracil
ASE	Allele-Specific Expression
COSMIC	Catalogue of Somatic Mutations in Cancer
DNA-seq	DNA sequencing
eQTL	Expression Quantitative Trait Loci
GCB	Gemcitabine
LINE	Long Interspersed Element
LoF	Loss of Function
RNA-seq	RNA sequencing
SINE	Short Interspersed Element
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TE	Transposable Element
TSG	Tumor Suppressor Gene

SUMMARY

Dysregulation of gene expression is a hallmark of cancer. Broadly speaking, my research is focused on the changes in gene expression that characterize the transition from normal to cancerous states, i.e. tumorigenesis. To study such changes, I performed integrated analysis of next generation sequencing data for matched normal and primary tumor samples from hundreds of patients across numerous different cancer types. By analyzing this sequencing data, I have been able to explore the global landscape of transcriptional reprogramming in cancer and discover how changes in the regulation of gene expression may be implicated in tumorigenesis. My thesis is focused on four specific areas of transcriptional reprogramming in cancer: (1) changes in the expression and activity of transposable elements (TEs), (2) changes in alternative splicing induced by TEs, (3) allele-specific expression of tumor suppressor genes (TSGs), and (4) gene expression changes that are implicated in cancer drug response.

TEs are known to be uniformly overexpressed in cancer, suggesting a possible role for their activity in tumorigenesis. I discovered a class of long interspersed nuclear elements (the LINE-1 family) with elevated levels of expression and activity in three different cancer types, and I showed examples where cancer-specific LINE-1 insertions disrupt enhancers, leading to the down-regulation of TSGs.

TEs are also implicated in the creation of novel splicing isoforms, and aberrant alternative splicing has been associated with tumorigenesis for a number of different cancers. Integrated analysis of genome sequence and transcriptome data revealed thousands of TE-generated alternative splice events genome-wide, including close to 5,000

events distributed among cancer associated genes. I explored the functional implications of specific cases of isoform switching, whereby TE-induced isoforms of cancer associated genes show elevated levels of relative expression in tumor samples.

A closer look at TSG expression in matched normal and tumor samples indicated that functionally important changes in patterns of allele-specific expression in individuals heterozygous for loss-of-function TSG alleles is a significant factor in cancer onset/progression. These results identified a variety of molecular mechanisms that contribute to the observed changes in allele-specific expression patterns in cancer with allele-specific alternative splicing mediated by anti-sense RNA emerging as a predominant factor. Furthermore, analysis of the genomic variation for world-wide human populations demonstrates that loss-of-function TSG alleles are segregating at remarkedly high frequencies implying that a significant fraction of otherwise healthy individuals may be pre-disposed to developing cancer.

For the final study of my thesis research, I applied the gene expression data from primary tumor samples to build predictive models of cancer drug response for two common chemotherapeutics: 5-Fluorouracil and Gemcitabine. My gene expression based models predict whether patients will respond to individual therapies with up to 86% accuracy. The genes that I found to be most informative for predicting drug response were enriched in well-known cancer signaling pathways highlighting their potential significance in prognosis of chemotherapy.

Research Advance 1: Patterns of transposable element (TE) expression and TE insertion activity in matched normal and primary tumor samples were analyzed for three

cancer types: breast invasive carcinoma, head and neck squamous cell carcinoma, and lung adenocarcinoma. We found high levels of somatic TE activity in normal and cancer samples across these diverse tissue types. We also observe a consistent increase in L1 transcript expression and L1 insertional activity in primary tumor samples for all three cancer types. Finally, we were able to investigate specific cases of putative cancer-causing TE mutations in further detail using genome feature analysis. These results inform the TE research community about unexpectedly high levels of somatic TE activity and a uniform increase in L1 expression and transposition across diverse cancer tissue types.

Research Advance 2: This chapter broadly characterizes the role of human TEs in generating alternatively spliced isoforms in cancer. To do so, we screened for the presence of TE-derived sequences co-located with alternative splice sites that are differentially utilized in paired normal versus cancer tissues. We analyzed a comprehensive set of alternative splice variants from 614 matched normal-tumor tissue pairs characterized via RNA-seq as part of The Cancer Genome Atlas (TCGA). Our algorithm uncovered close to 5,000 TE-generated alternative splice events distributed among Catalogue Of Somatic Mutations In Cancer (COSMIC) census genes that have been causally implicated in cancer. SINEs and LINEs were found to contribute the majority of TE-generated alternative splice sites in cancer genes. Differential expression analysis was used to detect TE-derived splicing events that are over-expressed in cancer tissues. A number of cancer-associated genes – *MYH11, WHSC1,* and *CANT1* – were shown to have overexpressed TE-generated isoforms across a range of cancer types.

Research Advance 3: Cancer has long thought to be a disease which develops from *de novo* cancer driver mutations in oncogenes and tumor suppressor genes (TSGs). The

purpose of this chapter is to examine how TSGs contribute to tumorigenesis. Accordingly, loss-of-function (LoF) mutations in TSGs were analyzed within 2504 individuals from 1000 Genomes Project (1KGP) and 233 patients across four diverse cancer types from The Cancer Genome Atlas (TCGA). A large fraction of 1KGP individuals were identified as carriers of heterozygous LoF mutations in TSGs. However, compound heterozygosity of LoF mutations in at least one TSG was only found in 20% of our TCGA patients. Further, analysis of allele specific expression (ASE) in these tumors identified several TSGs where the mutant allele is overexpressed relative to the reference allele. This evidence of ASE suggests TSGs have the potential to drive tumorigenesis in the heterozygous condition, if the reference allele is sufficiently repressed. A variety of molecular mechanisms contributing to ASE were identified including allele-specific alternative splicing induced by anti-sense RNA.

Research Advance 4: Both clinical and gene expression data from TCGA primary tumor biopsies were used to build models that predict patient response to cancer drugs. Our research focused on two common chemotherapeutics, Fluorouracil and Gemcitabine, and developed models with prediction accuracy up to 86%. These models will provide much needed decision support for oncologists when selecting second-line therapies. Traditionally, when oncologists are faced with this decision, they have little to no information about which drug will perform best for an individual patient. Therefore, our models, which create personalized predictions of response, provide essential information for these clinicians.

CHAPTER 1. INTRODUCTION

1.1 Transposable Elements (TEs)

Transposable elements (TEs) are DNA sequences, typically repetitive, that are capable of replicating and moving themselves in the host genome. TEs or "jumping genes" were first discovered in 1950 by Barbara McClintock while she was studying the maize genome at Cold Spring Harbor Laboratory in New York [1]. Eukaryotic TEs are broken down into two distinct classes based on the mechanism by which they transpose [2]. Class I TEs, or retrotransposons, require reverse transcription in order to transpose, following a "copy-and-paste" mechanism. Class II TEs, or DNA transposons, "cut-and-paste" simply moving their location instead of replicating.

1.1.1 L1 Insertions

Recent estimates suggest that close to 50% of the human genome is derived from TEs [3]. While most of these TEs are ancient remnants and no longer active, three families of TEs continue to retrotranspose in the human genome: Alu, L1 and SVA [4]. Of the three, L1s have been conclusively shown to be uniformly over expressed in several cancer types [5-9]. Unlike single nucleotide changes, L1 insertions are potentially far more disruptive and deleterious to the host genome. Consequently, their retrotransposition in the cell is normally suppressed via numerous mechanisms targeting different stages of the retrotransposition. These cellular inhibitory mechanisms include RNA degradation, autophagy-signaling pathways, inhibition of RNP formation and/or localization to stress

granules Each of these mechanisms/pathways is controlled by a set of genes that act as multiple failsafes. Failure in some (or all) of these machineries can lead to deleterious L1 insertions. Other mechanistic failures that can lead to the increase of L1 insertions in the genome include epigenetic changes such hypomethylation (or failure of maintaining methylated L1).

L1 insertions may contribute to tumorigenesis if the insertion causes genomic instability (improper chromosomal pairing during mitosis and meiosis), large genomic deletions, or disruption of cancer driver genes by inserting nearby (Figure 1).

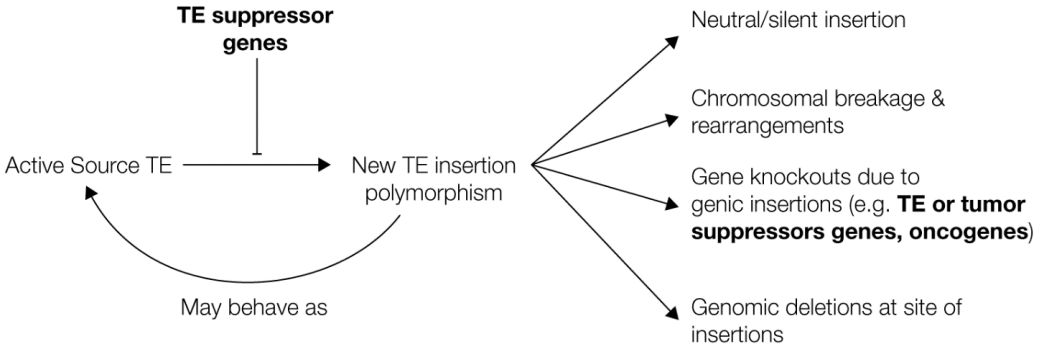

Figure 1 Contribution of TE insertions to tumorigenesis.

1.2 Alternative Splicing

Recent studies estimate there are over 21,000 human protein-coding genes [10, 11]. DNA is transcribed into RNA where it may undergo a process called alternative splicing. Alternative splicing is an editing process that involves removing intronic regions, leaving only protein-coding regions (*i.e.* exons) to make up the final processed messenger RNA

(mRNA). Through this process a single gene can code for multiple proteins. It is estimated that there are ~100,000 human proteins resulting from just ~21,000 genes.

There are five main types of alternative splicing: 1) alternative 3' splice site, 2) alternative 5' splice site, 3) intron retention, 4) exon skipping, and 5) mutually exclusive exons. All of these alternative-splicing events contribute to transcriptome and proteome diversity, significantly affecting the function of molecular processes that could contribute to disease states such as cancer. In fact aberrant alternative splicing has been associated with many cancers [12, 13].

The mechanism of alternative splicing has been attributed mainly to *cis*-acting regulatory elements in the mRNA sequence. These elements determine which exons are spliced by binding splicing facilitator proteins that act in *trans* to suppress splicing. Splicing inhibitors, splicing silencers and splicing activators all have a role in determining the location and ability of the spliceosome to assemble. It has also been shown that antisense RNAs play a vital role in alternative splicing. An antisense transcription-mediated mechanism of splicing has been described in humans where matching sense and antisense transcripts form double-stranded RNA leading to splice site masking [14]. Furthermore, antisense oligonucleotides have been designed to interfere with splice sites and successfully modulate splicing in a therapeutic manner [15-17].

1.3 Tumor Suppressor Genes (TSGs)

Cancer is thought to arise from mutations in genes that control cellular proliferation, differentiation and homeostasis [18]. These cancer-associated genes can be classified in two categories: oncogenes and tumor-suppressor genes (TSGs). Jointly, over expression of oncogenes and loss of TSG function drive tumorigenesis. Loss-of-function (LoF) mutations in TSGs can be confidently characterized with existing genomic techniques. This enables us to study the impact of these mutations on tumorigenesis. LoF mutations affecting TSGs frequently act in a recessive manner - and therefore must occur in both alleles of a TSG for a cell to become cancerous. This idea, known as the "two-hit" hypothesis, was formulated by geneticist Alfred Knudson in 1971 [19]. While studying retinoblastoma, a rare form of childhood cancer where tumor formation occurs in the back of the eye, Knudson looked at 48 patients and logged age, family history, and whether the disease was unilateral or bilateral. Using mathematics, Knudson was able to determine that the data followed a two-hit model as well as differences between hereditary and non-hereditary groups. Given that LoF, by mutation or deletion, of TSGs is common in cancer, learning more about them is critical for understanding carcinogenesis.

1.4 Allele-Specific Expression (ASE)

Humans are diploid organisms and as such inherit two copies of each gene – one maternal and one paternal. Allele-specific expression refers to the phenomena where one of these copies of a gene, or allele, is expressed significantly higher or lower than the other. Imprinted genes are examples of extreme ASE or monoallelic expression.

Allele-specific expression analysis is a powerful method for discovering *cis*-regulatory variability by comparing expression levels of reference (wild type) and alternative (mutant) alleles at all heterozygous variant positions in the exome [20]. Analyzing ASE in matched normal and tumor samples will uncover changes in quantities or patterns of ASE between samples.

There are various reasons why the expression of alleles may vary. One explanation is there is an imbalance of the alleles in the DNA. This could be due to tumor heterogeneity; since tumors are polyclonal, some single nucelotide polymorphisms (SNPs) may be specific to a clone in high or low frequency. It may also be attributed to DNA copy number variations (CNVs), which have been shown to be widespread in cancer [21]. Another possibility includes regulatory mutations. *Cis*-acting mutations such as expression quantitation trait loci (eQTLs) in promoter or enhancer regions may alter the expression of just one allele. Other scenarios include environmental factors that silence the maternal or paternal allele. Imprinter genes are a fine example of this, as is X-chromosome inactivation, whereby the inactive allele is packaged as heterochromatin such that it is transcriptionally inactive. Global DNA hypomethylation and tumor suppressor hypermethylation are two epigenetic alterations associated with many cancers that could play a role in allele-specific expression.

1.5 Precision Oncology

Precision oncology revolves around the idea that every patient's cancer is different and thus each patient should be treated differently. There are a number of diverse strategies in

cancer medicine that can be classified as precision oncology. These methods range from "targeted" approaches – whereby specific genes that are over expressed in the tumor are identified and reversed – to more modern-day approaches such as utilizing next-generation sequencing data to guide therapies. While traditional approaches have relied on knowledge of molecular pathways, (*i.e*. "cause and effect") [22], it's become increasingly apparent that our understanding of such processes is still limited [23].

One alternative approach to selecting therapies based on casual inference is guiding treatment decision through significant correlations in data. This is becoming more possible with the continued generation of big data in genomics as well as advances in one branch of artificial intelligence called machine learning. Toward that end, several machine-learning algorithms have been adopted for predicting cancer drug response including logistic regression, support vector machine, and random forest [24-26].

CHAPTER 2. PATTERNS OF TRANSPOSABLE ELEMENT EXPRESSION AND INSERTION IN CANCER

2.1 Abstract

Human transposable element (TE) activity in somatic tissues causes mutations that can contribute to tumorigenesis. Indeed, TE insertion mutations have been implicated in the etiology of a number of different cancer types. Nevertheless, the full extent of somatic TE activity, along with its relationship to tumorigenesis, have yet to be fully explored. Recent developments in bioinformatics software make it possible to analyze TE expression levels and TE insertional activity directly from transcriptome (RNA-seq) and whole genome (DNA-seq) next-generation sequence data. We applied these new sequence analysis techniques to matched normal and primary tumor patient samples from the Cancer Genome Atlas (TCGA) in order to analyze the patterns of TE expression and insertion for three cancer types: breast invasive carcinoma, head and neck squamous cell carcinoma, and lung adenocarcinoma. Our analysis focused on the three most abundant families of active human TEs: Alu, SVA and L1. We found evidence for high levels of somatic TE activity for these three families in normal and cancer samples across diverse tissue types. Abundant transcripts for all three TE families were detected in both normal and cancer tissues along with an average of ~80 unique TE insertions per individual patient/tissue. We observed an increase in L1 transcript expression and L1 insertional activity in primary tumor samples for all three cancer types. Tumor-specific TE insertions are enriched for private mutations, consistent with a potentially causal role in tumorigenesis. We used genome feature analysis to investigate two specific cases of putative cancer-causing TE

mutations in further detail. An Alu insertion in an upstream enhancer of the *CBL* tumor suppressor gene is associated with down-regulation of the gene in a single breast cancer patient, and an L1 insertion in the first exon of the *BAALC* gene also disrupts its expression in head and neck squamous cell carcinoma. Our results are consistent with widespread somatic activity of human TEs leading to numerous insertion mutations that can contribute to tumorigenesis in a variety of tissues.

2.2 Introduction

More than 50% of the human genome sequence is derived from transposable element (TE) insertions [3, 27]. The vast majority of TE-derived sequences in the human genome correspond to relatively ancient insertions that are no longer capable of transposition [4]. However, there are several families of human TEs that remain active to this day. The most abundant families of active TEs in the human genome are the Alu and SVA short interspersed nuclear elements (SINEs) along with the L1 Long Interspersed Nuclear Element (LINE) family [28-33]. Alu and SVA SINEs are non-autonomous TEs that are mobilized via the transpositional machinery encoded by the autonomous L1 family of LINEs. Recent evidence indicates that a handful of HERV-K endogenous retroviral elements also remain active in the human genome [34].

Active TE families are of great interest since they have the ability to generate de novo mutations, many of which have been linked to human disease [35, 36]. For instance, TE insertions have been shown to contribute to the etiology of a variety of different cancer types [37, 38]. Numerous recent studies have used a combination of next-generation

sequence analysis, followed by validation with PCR and/or Sanger sequencing, to elucidate connections between TE activity and cancer [35, 39-42]. L1 insertions in particular have been implicated as potential cancer causing mutations in those and other studies [6-9, 43]. L1 activity is thought to promote tumor development by causing genomic instability, via impaired chromosomal pairing during mitosis, and/or by disrupting coding or regulatory sequences [44].

Many of the studies that have related TEs to cancer have considered TE expression, at the transcript or protein level, and TE insertional activity separately. A number of different cancer types are positive for L1 transcript expression [45], and L1 proteins have been shown to be ubiquitously expressed in both normal and tumor samples from the same individuals [42, 46-50]. There is also evidence suggesting that L1 protein expression can be limited to tumor tissues and thereby serve as a useful cancer biomarker; nearly half of all human cancers are exclusively immunoreactive for L1-ORF1 encoded proteins [51]. The expression of L1 proteins in tumors has been shown to affect the expression of a number of cancer-related genes, including the down-regulation of tumor suppressors [52]. With respect to TE insertional activity, studies on matched normal and tumor tissues have found that novel L1 insertions occur at high frequencies in lung cancer genomes (Iskow et al., 2010). Such insertions frequently occur in oncogenes and tumor suppressors, underscoring their putative role in tumorigenesis (Lee et al., 2012).

A principal challenge when interpreting cancer genomes is distinguishing between so-called passenger and driver mutations. While passenger mutations are present in cancer genomes, they are not considered to contribute to cancer progression; instead, they are simply somatic mutations that arise during carcinogenesis and are carried along during

clonal expansion. Driver mutations, on the other hand, are causal mutations that are directly implicated in carcinogenesis and the promotion of cancer growth [53-55]. To date, only a few studies have directly implicated TE insertions as cancer driver mutations. One such study analyzed 19 hepatocellular carcinoma genomes utilizing the RC-Seq methodology [56] and discovered two separate L1 insertions that initiate tumorigenesis via distinct oncogenic pathways [41]. This study found L1 insertions in two different tumor suppressor genes: Mutated in Colorectal Cancers (MCC) and Suppression of Tumorigenicity (ST18). Most recently, a role for L1 insertional activity was conclusively demonstrated for colorectal cancer caused by an insertion in the APC tumor suppressor gene [43]. This paper describes a somatic L1 insertion into one copy of the APC gene that, when coupled with a point mutation in the other copy of the gene, initiates tumorigenesis through the two hit colorectal cancer pathway.

Owing to parallel developments in genomics and bioinformatics, it is now possible to jointly analyze the patterns of TE transcript expression and TE insertional activity in human cancers. The Cancer Genome Atlas (TCGA) provides access to both transcriptome sequence data (RNA-seq) and whole genome sequence data (DNA-seq) for a number of matched normal and primary tumor sample pairs from individual patients [57]. In addition, recently developed bioinformatics algorithms allow for the detection of TE transcripts directly from RNA-seq data [58] as well as for the characterization of novel TE insertions from DNA-seq data [59, 60]. We took advantage of these developments in order to evaluate the patterns of both TE expression and insertional activity in three cancer types: breast invasive carcinoma, head and neck squamous cell carcinoma, and lung adenocarcinoma (Figure 2 and Figure 21). We observed a simultaneous increase of L1

transcript expression and L1 insertional activity for primary tumor samples for all three cancers, and we evaluate individual cases of TE insertions that are implicated as potential cancer-causing mutations.

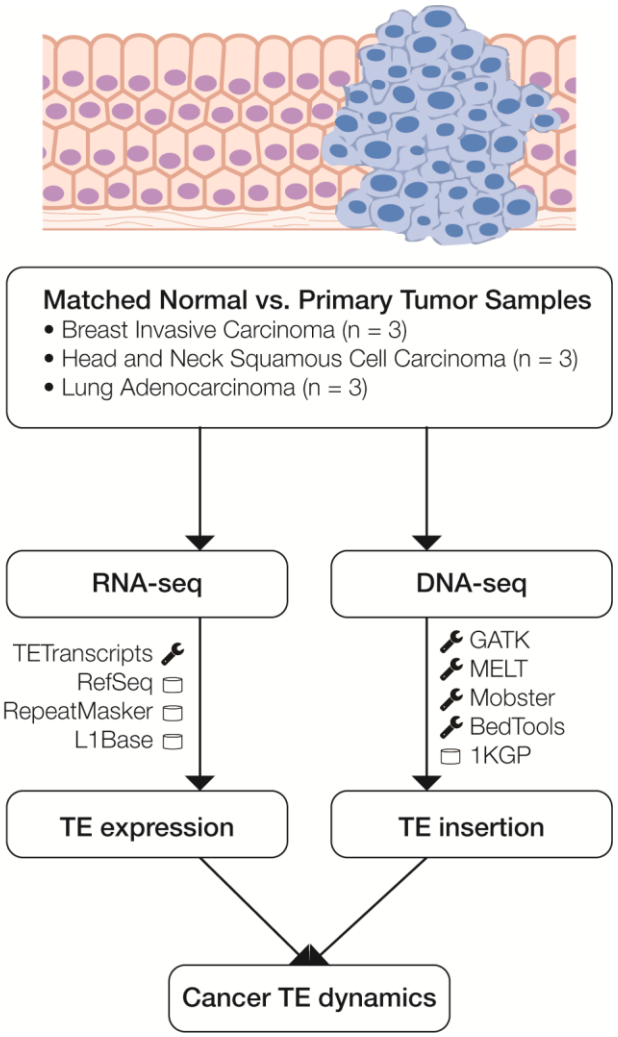

Figure 2 Scheme of the analytical design used in this study.

Matched normal and primary tumor samples for three cancer types were analyzed using transcriptome (RNA-seq) and whole genome (DNA-seq) data. RNA-seq data was analyzed to compare normal versus cancer expression levels, and DNA-seq data was analyzed to identify somatic TE insertion events. The main bioinformatics programs (wrench) and databases (cylinder) used for each phase of the analysis are indicated.

2.3 Materials and Methods

2.3.1 Genome and Transcriptome Sequence Data

Whole genome sequence data (DNA-seq), transcriptome sequence data (RNA-seq) and patient metadata for matched normal and primary tumor tissue samples from nine cancer patients were acquired from The Cancer Genome Atlas (TCGA) [57] via the Cancer Genomics Hub (CGHub) using the download client GeneTorrent [61]. The nine participants included three breast invasive carcinoma patients, three head and neck squamous cell carcinoma patients and three lung adenocarcinoma patients (Table 1). DNA-seq and RNA-seq data were accessed as BAM files of paired-end Illumina sequence data aligned against the human genome reference sequence (build hg19). BAM files containing sequence alignments were validated for quality using FASTQC [62], and autosomes were extracted from the BAM files for downstream analysis using SAMtools [63].

Table 1 TCGA whole genome (DNA-seq) and transcriptome (RNA-seq) data sources for the patients analyzed in this study.

ID	TCGA Barcode	Cancer Type	Sex	Age	Sample Type[a]	Seq Depth	Read Len.
Breast 1	TCGA-BH-A0B3-11B-21D-A128-09	Breast Invasive Carcinoma	F	53	NT-W	42.4	100
	TCGA-BH-A0B3-11B-21R-A089-07				NT-R	5.5	50
	TCGA-BH-A0B3-01A-11D-A128-09				TP-W	40.2	100
	TCGA-BH-A0B3-01B-21R-A089-07				TP-R	5.4	50
Breast 2	TCGA-BH-A0BW-11A-12D-A314-09		F	71	NT-W	54.1	100
	TCGA-BH-A0BW-11A-12R-A115-07				NT-R	7	50
	TCGA-BH-A0BW-01A-11D-A10Y-09				TP-W	46.1	100
	TCGA-BH-A0BW-01A-12R-A115-07				TP-R	7.3	50
Breast 3	TCGA-BH-A0DT-11A-12D-A12B-09		F	41	NT-W	63.3	100
	TCGA-BH-A0DT-11A-12R-A12D-07				NT-R	7.7	50
	TCGA-BH-A0DT-01A-21D-A12B-09				TP-W	79.9	100
	TCGA-BH-A0DT-01A-21R-A12D-07				TP-R	6.6	50

Head 1	TCGA-CV-7255-11A-01D-2276-10	Head and Neck Squamous Cell Carcinoma	F	32	NT-W	6.9	101
	TCGA-CV-7255-11A-01R-2016-07				NT-R	7.5	48
	TCGA-CV-7255-01A-11D-2276-10				TP-W	5.8	101
	TCGA-CV-7255-01A-11R-2016-07				TP-R	7.1	48
Head 2	TCGA-CV-7416-11A-01D-2334-08		F	29	NT-W	7.7	101
	TCGA-CV-7416-11A-01R-2081-07				NT-R	5.9	48
	TCGA-CV-7416-01A-11D-2334-08				TP-W	28.6	101
	TCGA-CV-7416-01A-11R-2081-07				TP-R	6	48
Head 3	TCGA-CV-6959-11A-01D-1911-02		M	48	NT-W	38.3	51
	TCGA-CV-6959-11A-01R-1915-07				NT-R	8.5	48
	TCGA-CV-6959-01A-11D-1911-02				TP-W	31.4	51
	TCGA-CV-6959-01A-11R-1915-07				TP-R	6.6	48
Lung 1	TCGA-44-6776-11A-01D-1853-02	Lung Adenocarcinoma	F	60	NT-W	38.9	51
	TCGA-44-6776-11A-01R-1858-07				NT-R	5.4	48
	TCGA-44-6776-01A-11D-1853-02				TP-W	6.9	51
	TCGA-44-6776-01A-11R-1858-07				TP-R	7.4	48
Lung 2	TCGA-50-5932-11A-01D-1753-08		M	75	NT-W	34.6	101
	TCGA-50-5932-11A-01R-1755-07				NT-R	4.2	48
	TCGA-50-5932-01A-11D-1753-08				TP-W	44.5	101
	TCGA-50-5932-01A-11R-1755-07				TP-R	7.4	48
Lung 3	TCGA-55-6984-11A-01D-1945-08		F	NA	NT-W	36.2	101
	TCGA-55-6984-11A-01R-1949-07				NT-R	4.9	48
	TCGA-55-6984-01A-11D-1945-08				TP-W	41	101
	TCGA-55-6984-01A-11R-1949-07				TP-R	5.2	48

[a] NT-D=Normal tissue DNA-seq, NT-R=Normal tissue RNA-seq, TP-D=Tumor primary DNA-seq, TP-R=Tumor primary RNA-seq

2.3.2 Gene and Transposable Element (TE) Expression Levels

Gene and TE expression levels were measured using RNA-seq data for the nine matched normal and primary tumor tissue samples. Gene expression levels were quantified as read counts mapped to NCBI RefSeq gene annotations [64]. TE expression levels – for Alu, L1 and SVA elements – were quantified using reads mapped to RepeatMasker annotations, which were subsequently analyzed with the TEtranscripts package [58]. The TEtranscripts program uses an expectation maximization (EM) algorithm to choose optimal unique TE locations for multi-mapped reads, thereby allowing for accurate

expression level measurements for active TE families. The TEtranscripts method was recently shown to yield more reliable measures of TE transcription levels compared to previously published methods, such as HTSeq-count, Cufflinks and RepEnrich [65-67]. The L1Base database was used to identify the genomic locations of 145 full length, intact elements from the most recently active L1 subfamily [68]. The set of full-length intact L1 sequences from the L1Base was generated by performing a BLAST search using the human genomic DNA sequences against the L1 template sequence [68]. L1Base was used to facilitate measures of active L1 element expression by limiting our analysis to RNA-seq reads that map to full-length, intact L1 sequences which retain the potential to be transpositionally active. This was done in an effort to ensure that the reads we analyzed were taken from potentially active L1 elements as opposed to older fixed elements, which could represent read-through transcripts initiated from nearby genomic promoters. The expression levels of these potentially active L1 elements were analyzed separately using the TEtranscripts method.

Differential expression levels between normal and cancer tissue pairs, for genes and TEs, were evaluated by comparing distributions of log10 transformed RNA-seq expression levels characterized as described above. The statistical significance levels of the observed differential expression between normal and cancer pairs were evaluated by comparing these distributions using the non-parametric Kolmogorov-Smirnov test. Statistical comparisons were done separately for each tissue (cancer) type: breast invasive carcinoma, head and neck squamous cell carcinoma and lung adenocarcinoma.

2.3.3 Transposable Element Insertion Detection

The genomic locations of novel TE insertions from matched normal and primary tumor tissue samples were predicted based on discordant read-pair mapping of DNA-seq data [39] (Table 2 Numbers of MELT and Mobster predicted TE insertions in matched normal (N) and primary tumor (T) samples across 9 individuals.. A scheme of our TE insertion detection analysis pipeline is shown in Figure 22. DNA-seq BAM files were realigned according to GATK's standard indel realignment method [69] to facilitate TE insertion detection. The programs MELT [59] and Mobster [60] were used together for TE insertion detection. These two programs were selected owing to their previously demonstrated superior performance for human TE insertion detection [70]. Only TE insertion sites that were found by both methods (i.e., the intersection of the predictions) were used for subsequent analysis. TE insertion predictions made by the individual programs were considered to represent the same insertion if they were found within ±100bp of each other. An additional filtering step was applied based on the number of mapped sequence reads (coverage) that support each TE insertion prediction. Only predictions with a minimum coverage of 5 reads and a maximum coverage of 4X the average sequencing depth of the sample were used for subsequent analysis. These upper and lower cut-off thresholds were empirically chosen based on the observed distributions of the numbers of discordant mapped read pairs used to call individual TE insertions. Read count distributions were computed individually for each program (MELT, Mobster) used and for each sample (Figure 23). The resulting distributions were typically bimodal with a lower peak (i.e., with lower read count support) that we considered to be enriched for potential false positive TE insertion calls. The lower cut-off threshold of 5 reads was chosen to minimize such false

positives, and the upper cut-off threshold was chosen to remove calls made in genomic regions that show anomalously high numbers of mapped reads, which tend to be enriched for ambiguously mapped reads.

The number of observed versus expected counts of unique L1 insertions were compared for matched normal and primary tumor tissue samples. The observed counts were taken from the TE detection pipeline, and the expected counts were computed as the ratio of unique insertions seen in matched normal versus primary tissue for all TEs multiplied by the total number of observed L1 insertions. The significance of the difference between the observed versus expected counts of unique L1 insertions was evaluated using the Fisher's exact test. Counts of TE insertions for matched normal and primary tumor tissue samples were characterized based on their frequencies from the 1000 Genomes Project (1KGP) [59] and grouped into three distinct frequency bins. The distributions of TE insertion counts across the three frequency bins were compared for matched normal and cancer samples for the different tissue types analyzed here, and the significance of the differences between these distributions were evaluated using the Kolmogorov-Smirnov test.

Table 2 Numbers of MELT and Mobster predicted TE insertions in matched normal (N) and primary tumor (T) samples across 9 individuals.

Participant ID	TE Insertions in Matched Normal Tissue				TE Insertions in Tumor Primary Tissue			
	Alu	SVA	L1	Total	Alu	SVA	L1	Total
Breast 1	913	28	127	**1,069**	853	33	110	**997**
Breast 2	1,004	21	121	**1,147**	1,160	54	143	**1,358**
Breast 3	1,012	63	139	**1,215**	952	60	136	**1,149**
Head 1	984	72	140	**1,197**	741	66	107	**915**
Head 2	945	25	131	**1,102**	832	26	138	**997**
Head 3	860	36	108	**1,005**	819	41	112	**973**
Lung 1	716	29	92	**838**	780	36	113	**930**
Lung 2	806	25	103	**935**	701	20	94	**816**

Lung 3	856	21	110	**988**	746	14	100	**861**

2.3.4 TE Insertion Genome Feature Analysis

The genomic locations of novel TE insertions were considered with respect to several genomic features using the BEDTools program [71]: RefSeq genes [64], COSMIC tumor suppressor genes [72], and enhancer elements defined by chromatin states [73]. The population allele frequencies of the predicted TE insertions were computed from the Phase 3 release of the 1KGP [59] as previously described [74].

2.4 Results and Discussion

2.4.1 TE Expression Levels in Matched Normal versus Primary Tumor Tissue Samples

RNA-seq data were used to evaluate the differences in TE expression levels between matched normal and primary tumor tissue samples as described in the Materials and Methods. The observed differences in gene expression levels between normal and tumor tissue were compared to differences in TE expression levels for breast invasive carcinoma, head and neck squamous cell carcinoma and lung adenocarcinoma. There are no significant differences observed for the distributions of gene expression levels between matched normal and primary tumor tissue pairs for any of the three cancer types analyzed here (Figure 3). Similarly, when all three families of potentially active TEs (Alu, L1 and SVA) are considered together, there is no significant difference seen for the overall levels of expression between matched normal and tumor tissue. However, when full-length,

potentially active L1 sequences are considered alone, we observe statistically significant increases in L1 expression levels for all three cancer types.

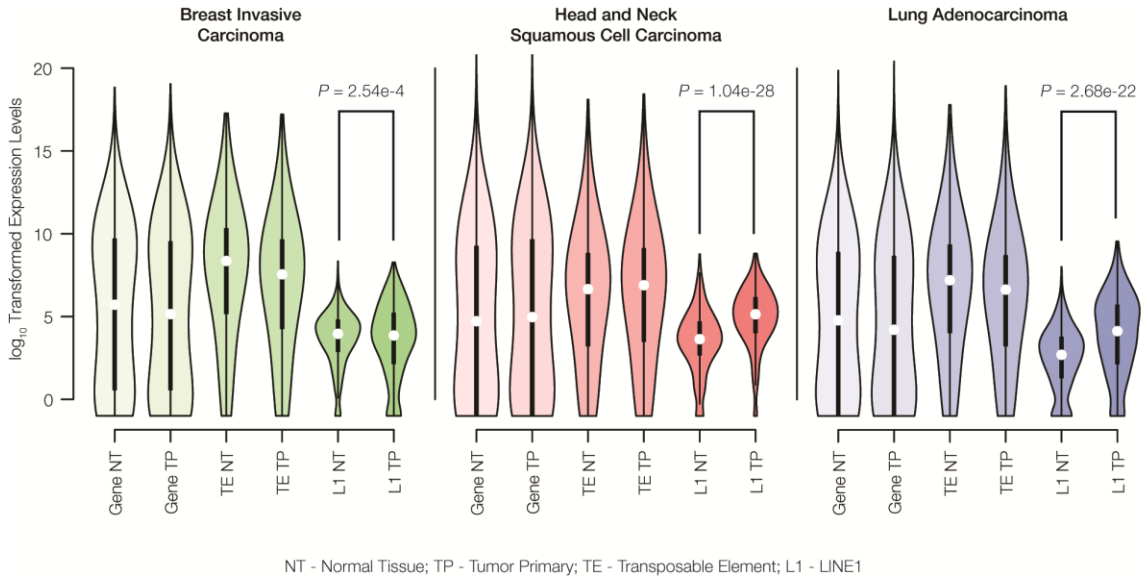

Figure 3 Gene expression dynamics for matched normal versus primary tumor tissue pairs.

Normal tissue (NT) and tumor primary (TP) expression levels were measured for genes, transposable elements (TEs) and LINE1 elements (L1s) via analysis of RNA-seq data as described in the Materials and Methods. Expression levels are shown as distributions of log_{10} transformed read counts, and normal versus tumor comparisons are shown for breast invasive carcinoma (green), head and neck squamous cell carcinoma (red) and lung adenocarcinoma (blue). For each tissue type, the significance levels of the differences in L1 expression between normal and cancer pairs are indicated with P-values from the Kolmogorov-Smirnov test.

The methods that we used to characterize TE expression levels include several analytical controls aimed to ensure that only genuine TE-initiated transcripts, from members of potentially active families, are measured. Nevertheless, the lack of a difference between normal and tumor expression levels observed when all three active TE families were considered together could reflect technical difficulties with identifying bona fide TE transcripts that are initiated from element promoters as opposed to TE sequences that are

passively expressed as part of longer genic transcripts. This is particularly true for Alu elements, many of which are found in the introns of human genes and transcribed as read-through transcripts initiated from RNA Pol II gene promoters [75]. Our confidence in the ability to measure L1-initiated transcripts is higher owing to the focus on previously identified full-length, intact elements that are located in intergenic regions. In any case, the up-regulation of L1s in cancer that we observed has potential implications for increased TE insertional activity for all three families, since L1 encoded proteins are responsible for the cis retrotransposition of L1s as well as the trans activation of Alu and SVA elements [76, 77]. We analyzed the same pairs of matched normal and primary tumor tissues to evaluate whether the observed increase in L1 expression corresponds to increased transpositional activity of human TEs.

2.4.2 Novel TE Insertions in Matched Normal and Primary Tumor Tissue Samples

It is now possible to characterize the genomic locations and copy numbers of individual TE insertions from whole genome DNA-seq data owing to recent developments in computational genomics software [39, 70]. This technological advance is exemplified by the recent Phase 3 release of the 1KGP, which includes a complete genome-wide census of polymorphic TE insertion sites for 2,504 individuals across 26 human populations [59]. We analyzed whole genome DNA-seq data using computational methods for TE insertion detection (see Materials and Methods) in order to compare TE insertional activity between matched normal versus primary tumor tissue samples.

When all three families of active human TEs are considered together, we observed a total of 3,672 TE insertions across the nine individuals analyzed for normal and cancer tissue pairs, 693 of which are unique insertions found in only one individual and one tissue type. In other words, we observe an average of ~77 unique somatic TE insertions per person, i.e., 'private' TE insertions. This estimate is similar to the value of ~90 unique (presumably germline) TE insertions that we previously observed for individuals from the 1KGP [74]. A large majority of the observed TE insertions – 81% for all TEs and 62% for L1s alone – are shared between the normal and tumor tissue types of an individual, suggesting that they represent germline insertions (Figure 4A). There are 1.3x more unique TE insertions seen for tumor compared to normal tissue, and this effect is more pronounced for L1s alone, which are 2x more abundant in tumor tissue samples. Accordingly, there is a statistically significant excess of observed versus expected L1 insertions in tumor versus normal tissue ($P = 0.019$) (Figure 4B). These results are consistent with a potential role for L1 transpositional activity in tumorigenesis for the cancer types analyzed here, as has been previously suggested for several different cancers (Morse et al., 1988; Iskow et al., 2010; Lee et al., 2012; Scott et al., 2016).

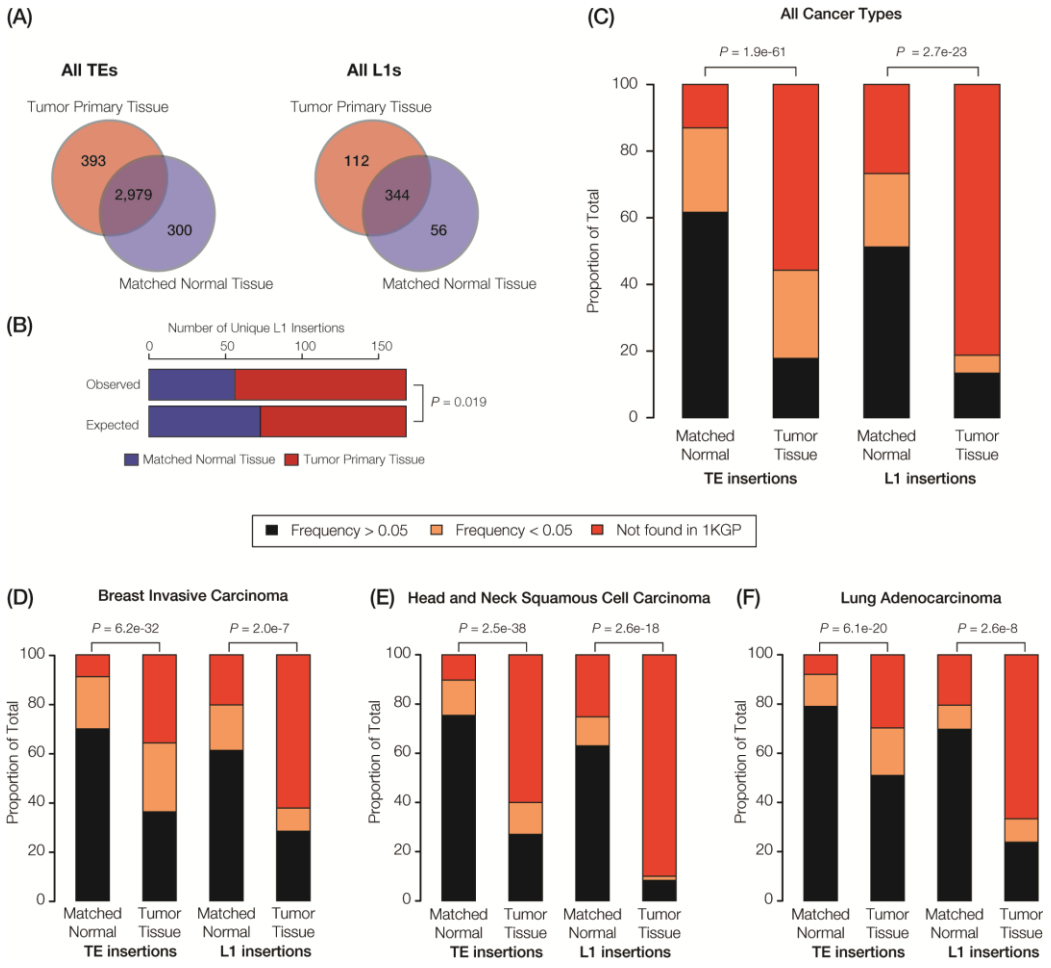

Figure 4 TE insertional activity in matched normal versus primary tumor tissue pairs.

Given the relatively high level of L1 insertional activity in the tumor tissue samples analyzed here, we tested whether tumor-specific L1 insertions are found at lower frequencies among the (presumably) healthy donors from the 1KGP compared to L1 insertions found in matched normal tissue. The idea was to evaluate whether the tumor-specific L1 insertions represent mutations that are private, and thereby more likely to be deleterious or disease-causing. To do this, individual TE insertions were classified as high frequency (>0.05), low frequency (<0.05) or private (absent) according to their previously characterized population (allele) frequencies from the 1KGP [59, 74].

When all three cancer types are considered together, there is a statistically significant excess of private and low frequency TE insertions observed for tumor compared to normal tissue (P = 1.9e-61) (Figure 4C). This effect is even more pronounced when L1 insertions are considered alone (P = 2.7e-23). The same pattern of an increased frequency of private L1 insertions in tumor tissue is observed (P < 2.0e-7) when all three cancer types are analyzed for sets of patients (Figure 4D-F) and when samples for individual patients are analyzed separately (Figure 24). The strongest effect is seen for head and neck squamous cell carcinoma. The pattern of a significant excess of private L1 insertions in tumor compared to normal tissue, observed for all three cancer types studied here, provides further evidence in support of a possible role for L1 activity in tumorigenesis.

It should be noted TE insertions found in low copy numbers may not be detectable using next-generation sequence analysis, whereas such insertions may be uncovered using more sensitive PCR-based approaches. False negatives of this kind will be more prevalent at low levels of sequence coverage. We have tried to control for this by using relatively high sequence coverage (~35X) studies here, but the conservative lower read count cut-off of 5 reads per TE insertion call that we used may still lead to missing TE insertion calls. Sequence based predictions can also yield false-positive TE insertion calls. In an effort to deal with this issue, we have only used high-confidence calls produced by two independent programs – MELT and Mobster – that we have recently shown to be most reliable for the detection of human TE insertions [70].

One other potential problem with the sequence based analysis relates to the base pair resolution with which TE insertions can be called via computational analysis of next-generation sequence data. Currently, the most accurate programs for calling TE insertions

from next-generation sequence data do not yet allow for the insertions to be precisely located to genomic regions at single base pair resolution. To account for this fact, TE insertions called within a window of ±100bp are considered to be co-located (Figure 22). It is possible that this approximation can lead to multiple TE insertion events being collapsed into a single event. Subsequent experimental confirmation of individual TE insertion calls of interest (e.g. potentially tumorigenic TE insertions) should help to provide certainty with respect to both their validity and their precise genomic locations.

2.4.3 Potentially Tumorigenic TE Insertions

Having established a potential role for transpositional activity in tumorigenesis using the genome-wide approaches described above, we wanted to search for specific examples where individual TE insertions could be implicated as possible cancer driver mutations. To do so, we performed an integrated analysis of TE insertion, gene expression and chromatin data (see Materials and Methods) in an effort to identify the cancer-specific TE insertions that are most likely to play a causal role in tumorigenesis. We considered TE insertions that are co-located with either exons or regulatory elements of previously characterized tumor suppressor genes to have the highest likelihood of being functionally relevant. We observed a total of 141 intragenic (35.9%) insertions and 246 intronic insertions (62.6%) out of the 393 total cancer-specific insertions in our dataset. None of these intergenic or intronic cancer-specific TE insertions were found to disrupt any known functional (regulatory) sequence element. Thus, consistent with previous studies, the vast majority of TE insertions that we observed are not likely to affect gene function or

expression in cancer. We did find 4 exonic TE insertions, along with 2 insertions located in regulatory elements, for known tumor suppressor genes (1.5% of the total). Here, we focus on two of these potential cases of cancer driver TE insertions, which could prove to be of interest to the TE and/or cancer research communities.

There is a private, breast cancer tumor-specific Alu insertion that is located within an upstream enhancer element that helps to regulate the expression of the Cbl Proto-Oncogene (CBL) gene (Figure 5A). CBL is classified as a tumor suppressor gene by the COSMIC database [72]. It has been found to be mutated or translocated in a number of cancers including acute myeloid leukemia [78, 79]; mutations in CBL are also the cause of Noonan syndrome-like disorder [80]. The CBL encoded protein functions as a negative regulator of signal transduction pathways [81], activation of which have been associated with cancer [82]. The tumor-specific Alu enhancer insertion that we characterized is associated with down-regulation of CBL expression, consistent with a potential role in tumorigenesis via the activation of signal transduction pathways associated with cell proliferation [82].

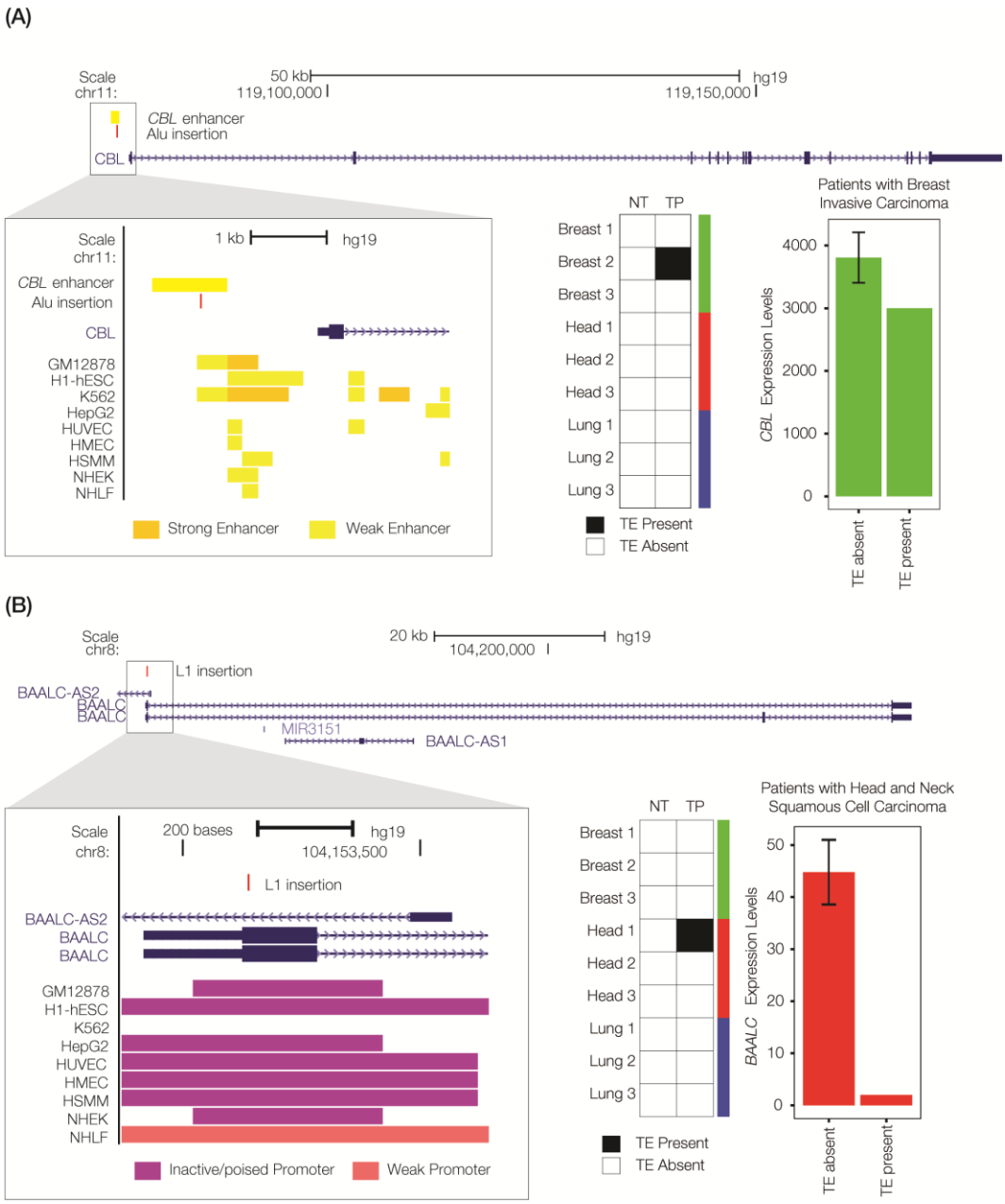

Figure 5 Private TE insertions implicated as potential cancer driver mutations.

We also found a private L1 insertion that was unique to a head and neck squamous cell carcinoma tissue sample, located within the first exon of the Brain and Acute Leukemia, Cytoplasmic (*BAALC*) gene (Figure 5B). As its name implies, the *BAALC* gene is expressed in the brain and related neural tissues, and it was first identified by association

with acute myeloid leukemia where it was shown to be overexpressed [83, 84]. TE insertions within exons are extremely rare and would presumably have a dramatic effect on gene function. Indeed, this particular insertion is associated with nearly complete inactivation of the *BAALC* gene. This is consistent with previous results showing that the presence of fixed L1 insertions genome-wide is strongly associated with the down-regulation of human gene expression [85]. A recent study has demonstrated that *BAALC* can inhibit extracellular signal-regulated kinase (ERK) mediated monocytic differentiation of AML cells [86]. Thus, down-regulation of *BAALC* would presumably result in a loss of control over cellular differentiation, consistent with a possible role in tumorigenesis. A recent study discovered a role for the change in methylation status of a cancer-specific L1 insertion in tumorigenesis (Scott et al., 2016); this could be an additional mechanism by which the *BAALC* L1 insertion observed here exerts a regulatory effect.

2.5 Conclusion

The results of our analysis show a surprisingly high level of somatic TE activity in the human genome. Abundant transcripts from members of all three active human TE families analyzed here – Alu, SVA and L1 – can be identified for both normal and cancer tissue samples. In addition, after filtering for high confidence TE insertion calls, we identified an average of close to 80 unique insertions for each tissue among the individual patients in our study. Thus, active human TE families retain the ability to transpose in somatic tissue thereby generating substantial levels of cellular heterogeneity among diverse tissues.

We also observe a correlated increase in both transcript expression levels and transpositional activity for L1 elements in cancer tissue samples when compared to matched normal tissue. Increased cancer expression of L1 elements is particularly relevant for TE insertional activity, since the L1 transpositional machinery is responsible for transposing non-autonomous Alu and SVA elements in trans along with L1 elements in cis. Our results are consistent with previous studies showing expression of L1 transcripts in lung cancer [45] and expression of L1 ORF1p in breast cancer [87], and tumor-specific L1 insertions have also previously been found in breast (Morse et al., 1988), head and neck (Helman et al., 2014), and lung tumors (Helman et al., 2014). We confirmed the presence of numerous tumor-specific L1 insertions in these three cancer types and identify two potentially tumorigenic TE insertions, an Alu insertion in the enhancer region of the tumor suppressor gene CBL and an L1 insertion in the first exon of the *BAALC* gene. These results underscore the potential for somatic TE activity to generate cellular heterogeneity and to contribute to the etiology of cancer across a wide range of human tissues.

CHAPTER 3. TRANSPOSABLE ELEMENT INDUCED ALTERNATIVE SPLICING IN CANCER

3.1 Abstract

Transposable element (TE) derived sequences comprise more than half of the human genome, and their presence has been documented to alter gene expression in a number of different ways, including the generation of alternatively spliced transcript isoforms. Alternative splicing has been associated with tumorigenesis for a number of different cancers. The objective of this study was to broadly characterize the role of human TEs in generating alternatively spliced transcript isoforms in cancer. To do so, we screened for the presence of TE-derived sequences co-located with alternative splice sites that are differentially utilized in normal versus cancer tissues. We analyzed a comprehensive set of alternative splice variants characterized for 614 matched normal-tumor tissue pairs across 13 cancer types, resulting in the discovery of 4,820 TE-generated alternative splice events distributed among 723 cancer-associated genes. SINEs (Alu) and LINEs (L1) were found to contribute the majority of TE-generated alternative splice sites in cancer genes. A number of cancer-associated genes – including *MYH11*, *WHSC1*, and *CANT1* – were shown to have overexpressed TE-induced isoforms across a range of cancer types. TE-induced isoforms were also linked to cancer-specific fusion transcripts, suggesting a novel mechanism for the generation of transcriptome diversity via trans-splicing mediated by dispersed TE repeats.

3.2 Background

Half or more of the human genome is derived from transposable element (TE) sequences, remnants of formerly mobile genetic elements that can replicate to extremely high copy numbers over time [88, 89]. TE sequences contribute to human gene regulation through a variety of distinct mechanisms [90-93]. Previous work from our own lab has documented the presence of TE-derived transcription factor binding sites [94-97], enhancers [98-102], chromatin insulators [103], microRNAs [104, 105], and anti-sense RNAs [106] along with TE-derived alternative transcription initiation [107-109] and termination sites [110].

The provisioning of alternative splice sites is another way that TEs can contribute to the complexity of the human transcriptome [111-113]. A role for TEs in alternative splicing of human genes was discovered via classic studies on Alu elements in the early 2000s. Investigators from the laboratories of Gil Ast and Dan Graur uncovered evidence of Alu-derived splice sites, as well as the inclusion of Alu elements in alternatively spliced exons, for a number of human genes [114, 115]. These studies suggested a potential role for TE-induced alternative splicing in disease, cancer in particular [116]. Nevertheless, compelling proof for such a connection has remained elusive.

The role of TEs in cancer has received substantial attention as of late [8, 37, 43, 117-119], and alternative splicing has itself been widely associated with tumorigenesis [12, 120-126]. As such, it seems reasonable to hypothesize that TE-induced alternative splicing could play an important role in cancer. Despite the seemingly obvious connections among – TEs, alternative splicing, and cancer – there has yet to be any systematic analysis on the

contribution of TEs to alternative splicing events in tumor tissue. The goals of this study were to (1) survey the global landscape of TE-induced alternative splicing across a variety of cancer types, and (2) identify individual cases where TE-derived splice sites are linked to splicing (isoform) alterations in cancer.

We analyzed 614 matched normal-tumor samples pairs for 13 cancer types, characterized as part of The Cancer Genome Atlas (TCGA). Integrated analysis of RNA-seq data and genome annotations were used to generate a genome-wide atlas of TE-derived alternative splice sites, and differential expression analysis of alternative splice variants was used to identify 'isoform switch' events, with TE-induced splice isoforms that show increased utilization in cancer samples. Our atlas of TE-induced alternative splice variants is made available to the research community via the UCSC Genome Browser. We go on to propose a potentially novel mechanism, whereby the dispersed repetitive nature of TE sequences facilitates the generation of fusion transcripts via trans-splicing events. Our TE trans-splicing mechanism is admittedly speculative at this time, and we suggest the kinds of tests that will need to be done to further interrogate our model.

3.3 Methods

A schematic overview of the analysis pipeline used for this study can be seen in Figure 1. A list of all data sources, programs, and statistical methods used in the study can be seen in Table 10.

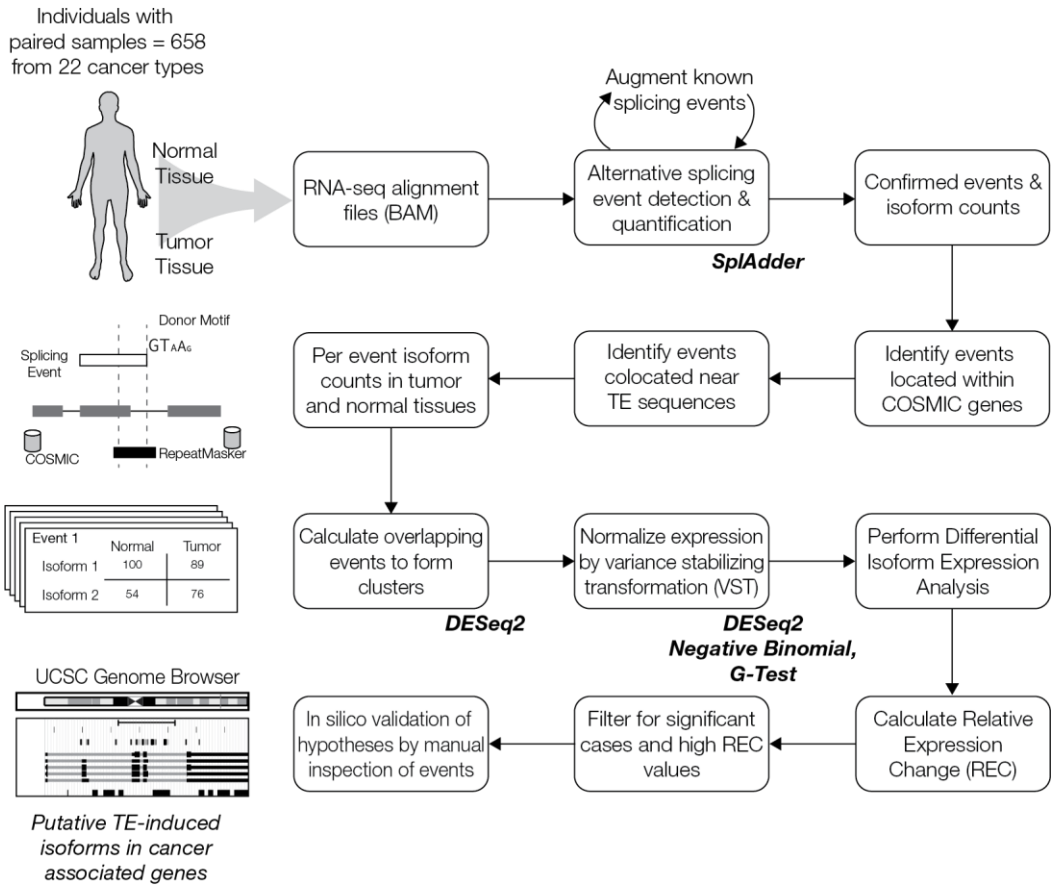

Figure 6 Bioinformatics analysis pipeline used for this study.

RNA-seq datasets from 658 paired normal-tumor TCGA samples from 22 cancer types were analyzed in this study. The schematic can be broadly divided into four stages: (row 1) detection of alternative splicing event and per-exon expression quantification, (row 2) identification of TE-derived alternative splicing events for cancer-associated genes, (row 3) statistical testing for differences in alternative splicing expression levels between normal and tumor tissues, and (row 4) evaluation of cases of interest to explore the potential functional impact of TE-derived alternative splicing on cancer.

3.3.1 Genomic Data

All analyses are based on the human genome reference sequence build hg19 (GRCh37). Genomic coordinates for NCBI RefSeq [127] and Ensembl transcript models, i.e. exon/intron boundaries, were taken from the UCSC Genome Browser [128]. Genomic coordinates for TE sequences were taken from the RepeatMasker annotations [129]. Overlap analysis of gene, TE, and alternative splice event coordinates were performed using the BEDTools program [71].

3.3.2 Alternative Splicing

The Catalogue Of Somatic Mutations In Cancer (COSMIC) Cancer Gene Census (CGC) was used to identify cancer-associated genes – oncogenes, tumor suppressor genes, and fusion genes – for subsequent alternative splicing analysis [130]. Transcriptome (RNA-seq) data for matched normal-tumor sample pairs of individual patients, across a variety of distinct cancer types, were taken from from The Cancer Genome Atlas (TCGA) for alternative splice site analysis (Figure 25). RNA-seq data were mapped to the human reference genome sequence and processed using the program SplAdder, as previously described [122], in order to characterize alternative splice events in cancer-associated genes. Four kinds of alternative splice events were analyzed here: intron retention, exon skipping, alternate 3' splicing, and alternate 5' splicing (Figure 26). For all observed alternative splice events, two distinct isoforms were defined and quantified. Isoform 1 and isoform 2 are operationally defined as the shorter and longer isoforms, respectively. Thus, isoform 1 corresponds to the TE-derived isoform for exon skipping, whereas isoform 2

corresponds to the TE-derived isoform for intron retention, alternate 3' splicing, and alternate 5' splicing.

Genomic coordinates for individual alternative splice sites and their corresponding isoforms are defined by the presence of overlapping RNA-seq reads for at least three individuals. Individual alternative splice events were characterized across all COSMIC genes, and each individual event was quantified as the number of reads mapping to the alternatively spliced exon. This was done for all genes from individual samples corresponding to each cancer type and its corresponding matched normal-tumor sample pair. Overlapping alternative splice event isoforms were clustered using single linkage clustering based on ≥75% overlap of splice site genomic coordinates, and cluster coordinates were defined as the minimum and maximum start and stop sites for the individual constituent splice sites. Alternative splice site cluster counts for all isoforms were calculated as the average counts across all individual constituent splice sites within any given tissue type.

3.3.3 Differential Expression (Splicing)

The program DESeq2 was used to normalize alternative splice site cluster counts using the variance stabilizing transformation (VST) technique [131]. Differential alternative splice isoform expression, between matched normal-tumor sample pairs, was measured using relative expression change (REC) and via a 2 x 2 contingency table with the G-test. For each alternative splice event, cluster average count values were computed across four conditions: (1) non-TE isoform normal, (2) TE isoform normal, (3) non-TE

isoform tumor, and (4) TE isoform tumor. The relative expression change (REC) value for individual alternative splice events are calculated as the normalized difference of the TE isoform in tumor versus normal tissue as described in Equation 1:

$$REC = \left(\frac{E_{Tumor}^{TE\ Isoform}}{E_{Tumor}^{non-TE\ Isoform} + E_{Tumor}^{TE\ Isoform}}\right) - \left(\frac{E_{Normal}^{TE\ Isoform}}{E_{Normal}^{non-TE\ Isoform} + E_{Normal}^{TE\ Isoform}}\right) \quad (1)$$

where, $E_{Normal}^{TE\ Isoform}$ is the average normalized cluster count for the TE-derived isoform across all individuals in normal tissue. The statistical significance of normal-tumor differential expression (splicing), i.e. differences in average alternative splice site cluster counts, was evaluated using a 2 x 2 contingency table with the G-test:

	Normal	*Tumor*	
Non-TE Isoform	$E_{Normal}^{non-TE\ Isoform}$	$E_{Tumor}^{non-TE\ Isoform}$	n_1
TE Isoform	$E_{Normal}^{TE\ Isoform}$	$E_{Tumor}^{TE\ Isoform}$	n_2
	n_N	n_T	N

3.3.4 Visualization

Individual cases TE-derived and differentially expressed alternative splice sites of interest were visualized using the UCSC Genome Browser. Locations of RNA-seq characterized alternative splice site clusters were compared to the locations of TE sequences and COSMIC gene exon/intron boundaries. Genomic coordinates of the TE-

induced alternatively spliced exons characterized here are distributed as a UCSC Genome Brower Track hub.

3.4 Results and Discussion

3.4.1 TE-derived Alternative Splice Sites and Cancer

We analyzed RNA-seq data for matched normal-tumor sample pairs from individual patients in order to characterize the genomic landscape of alternative splicing in cancer. A total of 678 patient samples among 22 different cancer types were considered for preliminary analysis; cancer types with <10 patient samples were subsequently excluded, yielding a final data set of 614 patients across 13 cancer types (Figure 25) (Table 3). We relied on a recently published approach to the characterization of alternative splicing in cancer, which has been shown to yield reliable results in terms of both characterizing and quantifying individual alternative splice sites and their corresponding isoforms [122]. We focused on four distinct types of alternative splicing events – intron retention, exon skipping, alterative 3' splicing, and alterative 5' splicing (Figure 26) – and modified the existing approach to yield tissue-specific counts of alternative splice site isoforms for individual patients (see Methods).

Table 3 TCGA patient samples analyzed in this study.

Cancer Type	TCGA Abbreviations	Number of Samples	Number of Participants
Breast invasive carcinoma	BRCA	220	110
Kidney renal clear cell carcinoma	KIRC	144	72
Thyroid carcinoma	THCA	116	58
Lung adenocarcinoma	LUAD	114	57

Prostate adenocarcinoma	PRAD	104	52
Liver hepatocellular carcinoma	LIHC	100	50
Lung squamous cell carcinoma	LUSC	98	49
Head and Neck squamous cell carcinoma	HNSC	84	42
Kidney renal papillary cell carcinoma	KIRP	62	31
Stomach adenocarcinoma	STAD	54	27
Colon adenocarcinoma	COAD	48	24
Kidney Chromophobe	KICH	46	23
Bladder Urothelial Carcinoma	BLCA	38	19

We then narrowed our analysis to a catalog of 723 known cancer-associated genes and focused on alternative splice sites in those genes that are derived from TE sequences. TE-derived splice sites were delineated by searching for canonical splice donor and acceptor site sequence motifs, located at 3' and 5' exon boundaries, that overlap with annotated TE sequences (Figure 27). Human TE sequences were divided into their four major classes – SINEs, LINEs, LTR, and DNA (Figure 28) – and the overall extent of their contribution to alternative splicing in cancer was evaluated. TE sequences contribute thousands of distinct alternative splice sites genome-wide, ranging from 10.5% of alternative 5' splice events to 14.0% of exon skipping events (Figure 29). TEs also contribute a substantial minority of the of alternative splice sites to cancer-associated genes. Across the 13 cancer types, TE-derived isoforms are a consistent minority, and the numbers of alternative splice sites are more similar for TE- versus non-TE-derived isoforms, compared to the relatively small differences seen for normal versus cancer samples (Figure 7A).

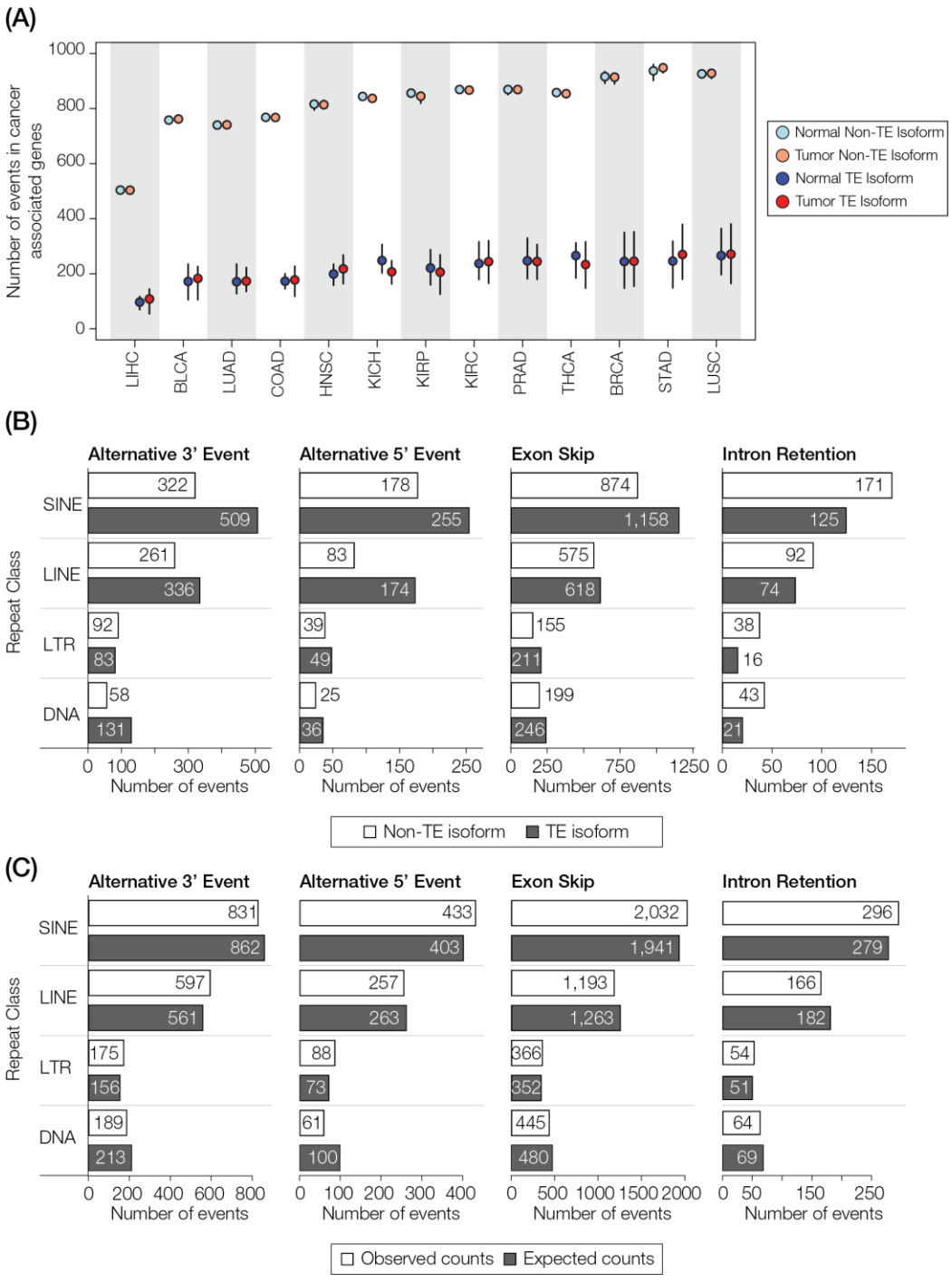

Figure 7 Overall landscape of TE-derived alternative splicing in cancer.

(A) Dot-and-whisker plot comparing the distribution of TE and non-TE isoforms in cancer-associated genes in normal (blue and light blue) and tumor (red and light-red) tissues across all samples within each cancer type. The median number of events are shown as dots and the outliers (defined classically as 1.5 x interquartile range) are shown as

whiskers. (B) Counts of the total number of unique TE and non-TE isoforms in cancer-associated genes is shown by the event type and TE class. (C) The observed counts of TE isoforms in cancer-associated genes for each event type and TE class is compared to expected counts.

At first glance, the overall landscape of TE-derived splicing isoforms in cancer-associated genes suggests the possibility that TE contributions to alternative splicing in cancer may not be very biologically significant. However, when alternative splicing events in cancer-associated genes are broken down by event type and TE-class, the potential contribution of TEs becomes more apparent. This is because for any given splice site where a TE is present, the numbers of TE-derived splice isoforms tend to outnumber the non-TE-derived isoforms (Figure 7B). This holds true for three out of the four alternative splice event types; for intron retention, the non-TE-derived isoforms are more common. Finally, it is interesting to note that there is no particular enrichment for the contributions of any given TE class to any of the four kinds alternative splice site events. The observed numbers of TEs from each class that contribute to these events are very similar to the expected numbers based on their background frequencies within cancer-associated genes (Figure 7C).

3.4.2 Differential Expression of TE-derived Splice Sites

We analyzed differences in the expression levels of alternative splice sites between matched normal-tumor sample pairs in an effort to evaluate the effects of individual TE-derived splice sites on cancer. The expression levels of individual alterative splice sites, and their corresponding isoforms, were quantified via normalized counts of mapped RNA-seq reads as detailed in the Methods section. For any given TE-derived splice site, there

are four possible expression counts for any individual patient: (1) non-TE isoform normal, (2) TE isoform normal, (3) non-TE isoform tumor, and (4) TE isoform tumor. Expression counts for these four conditions can be averaged across individuals to measure the relative expression changes (REC) of TE-derived isoforms in tumor compared to normal tissue and to evaluate the significance of this difference (Figure 30). Distributions of REC values for the four types of TE-derived splice sites across the 13 cancer types are shown in Figure 8. For the most part, these distributions are tightly clustered around the median value of 0, or no relative change, with sparsely populated tails that contain individual cases of potential interest. We evaluated a number of these outlier genes, with highly differentially expressed alternative splice isoforms in matched normal-cancer samples (Table 4), in an effort to explore potential functional implications of TE-derived splice sites in cancer.

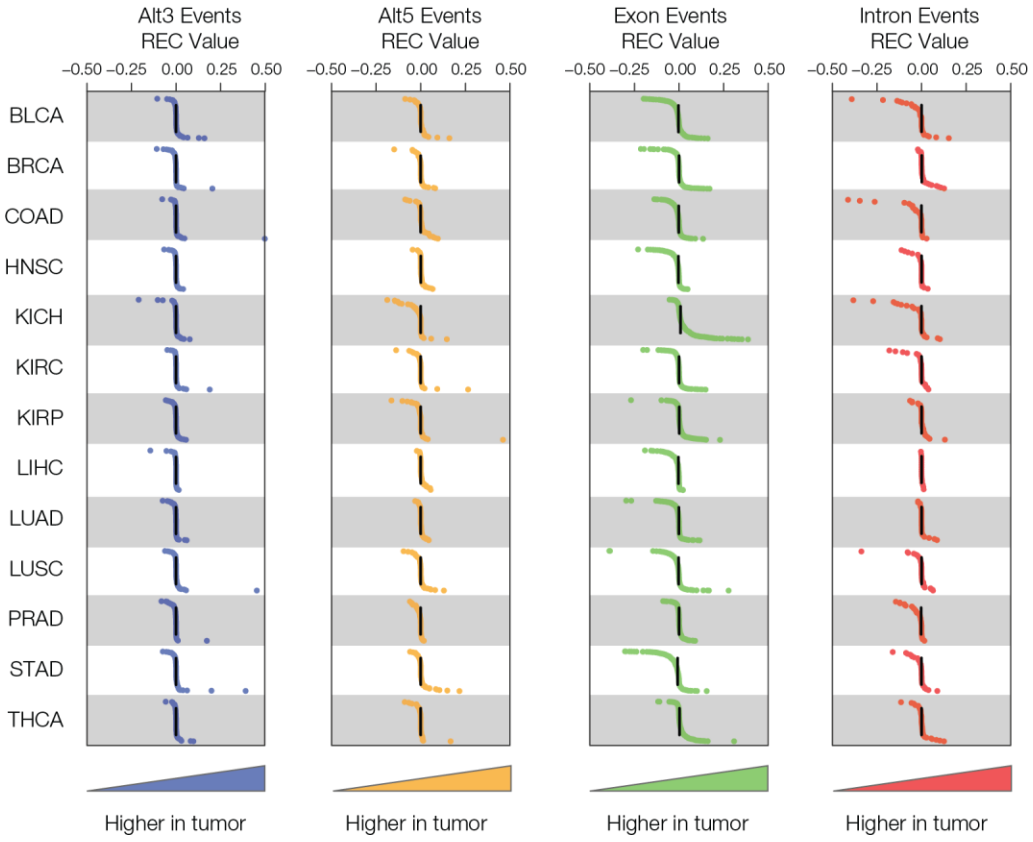

Figure 8 Differential expression of TE-derived alternative splice isoforms in tumor versus normal samples.

Distributions of the relative expression counts (REC) comparing TE-derived to non TE-derived alternative splice isoforms in tumor versus normal samples. The formula for REC is described in the Methods and in Figure 30. Data are shown for 13 cancer types and 4 alternative splice event types. Each dot represents an REC value derived from the average normalized expression counts of the TE- and non TE-derived isoforms in normal and cancer samples. Higher expression (counts) of the TE-derived isoform in tumor are shown on the right side of the panels, whereas lower expression is shown to the left.

3.4.3 Potential Functional Implications of TE-derived Splice Sites in Cancer

One particular result that stood out from this analysis was the observation that a few cancer-associated genes have extremely high counts of TE-derived alternative splicing events (Figure 9). The Kallikrein Related Peptidase 2 encoding gene KLK2 shows more

than twice as many TE-derived alternative splice sites compared to the second rank gene on this list. The KLK2 protein is primarily expressed in the prostate and has been shown to promote prostate cancer cell growth [132]. The connection between TE-derived alternative splicing and cancer is supported by the fact that all of the TE-derived isoforms observed here were identified in prostate adenocarcinoma samples. Alternative splicing of the KLK2 gene results in fusions with ETV1 and ETV4 in prostate cancer, and all of the known fusion transcripts for these genes are missing exon 3 of KLK2 [133, 134]. Interestingly, exon skipping events are by far the most abundant TE-derived splice isoforms seen for this gene (Figure 9). The large number of putative TE induced alternative splicing events in KLK2, specifically exon skipping, suggests TEs could play an important role in the manifestation of KLK2 fusion transcripts and their contribution to prostate cancer. Given their dispersed repetitive nature, it is possible that TE sequences serve as hotspots for the generation of fusion transcripts in cancer. We further explore this potential model for transcriptome diversification by TE sequences in the Conclusion section.

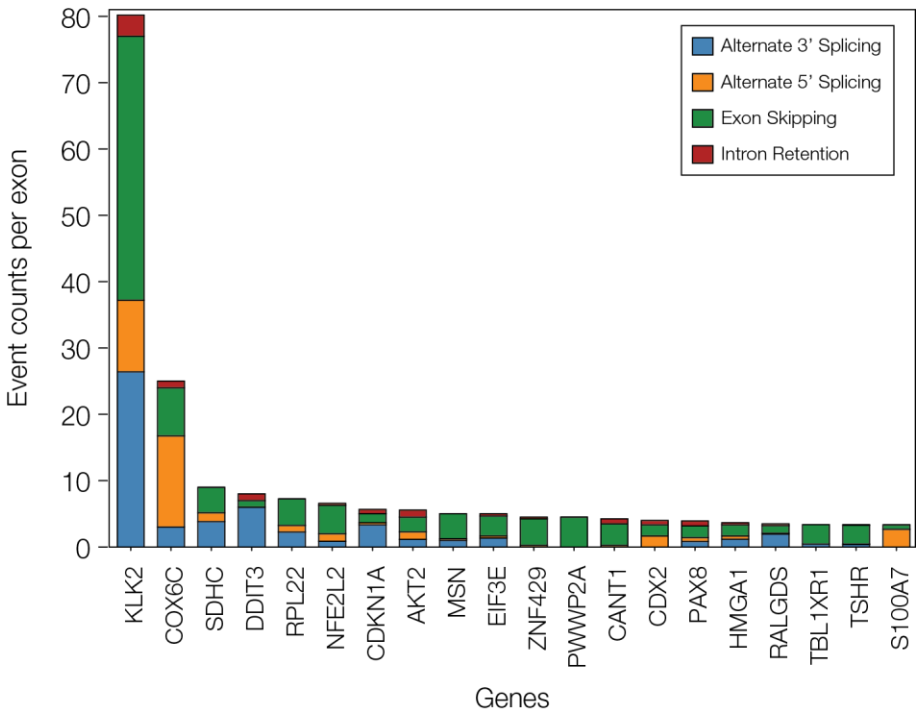

Figure 9 Frequency of TE-derived alternative splice events for individual genes.

The total numbers of alternative splice counts per exon are shown for each cancer-associated gene, broken down by the four alternative splice event types. Genes with the highest counts of TE-derived alternative splice events across all cancer types are shown.

Table 4 Top candidate TE-derived alternative transcription events.

Cluster	Gene	Events in Cluster	Cancer Type	%N_T	%T_T	Event Type
467	*MYH11*	2	Lung squamous cell carcinoma	6.5	51.8	Alt3
412	*CANT1*	2	Stomach adenocarcinoma	67.3	37.4	Exon
132	*WHSC1*	1	Stomach adenocarcinoma	73.1	42.8	Exon
412	*CANT1*	2	Breast invasive carcinoma	53.5	32.1	Exon
154	*KMT2D*	1	Stomach adenocarcinoma	21.3	31.8	Alt5
397	*POLG*	4	Stomach adenocarcinoma	34.8	54.6	Alt3
397	*POLG*	4	Bladder Urothelial Carcinoma	34.8	47.6	Alt3
261	*PML*	2	Kidney renal papillary cell carcinoma	57.2	70.3	Intron
261	*PML*	2	Breast invasive carcinoma	47.0	59.6	Intron
261	*PML*	2	Kidney Chromophobe	65.5	75.8	Intron

The Myosin Heavy Chain 11 gene *MYH11* encodes part of a hexameric protein that functions as a major contractile complex, converting chemical energy into mechanical energy through the hydrolysis of ATP. *MYH11* has been shown to contribute to tumorigenesis in both leukemia and non-small cell lung cancer (NSCLC) [135]. *MYH11* undergoes alternative splicing, yielding isoforms that are differentially expressed in tumor samples [136]. *MYH11* is also implicated in cancer-associated gene fusion events; for example, the *CBFB-MYH11* gene fusion plays an important role in leukemogenesis [137-139]. Here, we observe differential isoform expression of *MYH11* across 49 paired normal-tumor lung squamous cell carcinoma tissues, whereby a SINE (Alu) induces an alternative 3' splicing event that yields a longer version of exon 41 (Figure 10a). The longer SINE-derived isoform makes up 6.5% of the transcript population in normal samples compared to 51.8% in tumor samples (Figure 10b-c).

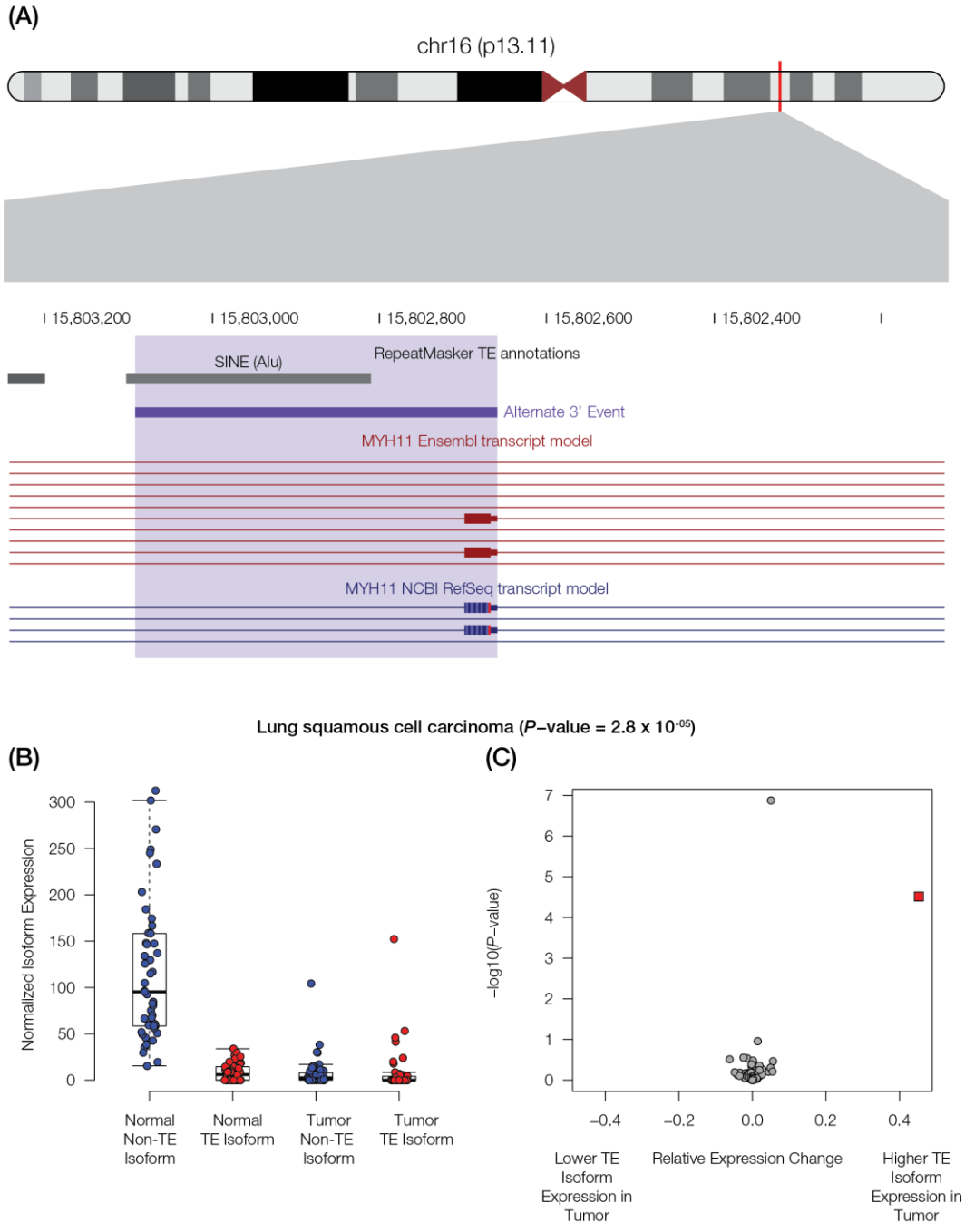

Figure 10 TE-derived alternative splicing in the *MYH11* gene.

(A) The location of MYH11 on the short arm of chromosome 16 is shown along with the specific location of its TE-derived alternative splicing event. A SINE (Alu) sequence provides an alternate 3' splice site resulting in an extended exon 41. (B) Distributions of the non-TE (blue) and TE-derived (red) isoforms are shown for matched normal (left) and lung squamous cell carcinoma samples (right). (C) Relative expression change (REC) values are plotted against the corresponding G-test P-values (see Methods and Figure 30)

for the matched normal and lung squamous cell carcinoma samples. The MYH11 TE-derived isoform values are shown as a red square.

The Wolf-Hirschhorn Syndrome Candidate 1 Protein gene *WHSC1*, also known as the Nuclear Receptor Binding SET Domain Protein 2 gene (NSD2), encodes a histone methyltransferase that catalyzes the dimethylation of histone 3 lysine 36 (H3K36). *WHSC1* expression is important for the epithelial-mesenchymal transition and metastasis in gastric cancer [140], and it is overexpressed in a number of different cancer types [141]. *WHSC1* has been shown to undergo complex alternative splicing. Most of the primary transcripts of *WHSC1* initiate from exon 3, which contains the proper translation initiation site, although a small fraction of transcripts retain upstream non-coding sequences including exons 1 and 2 [142]. Here, we identified a LINE (L1) element apparently responsible for an exon skipping event in exon 3, which occurs much more frequently in stomach adenocarcinoma primary tumor tissues (57%) when compared to matched normal tissues (27%) (Figure 11).

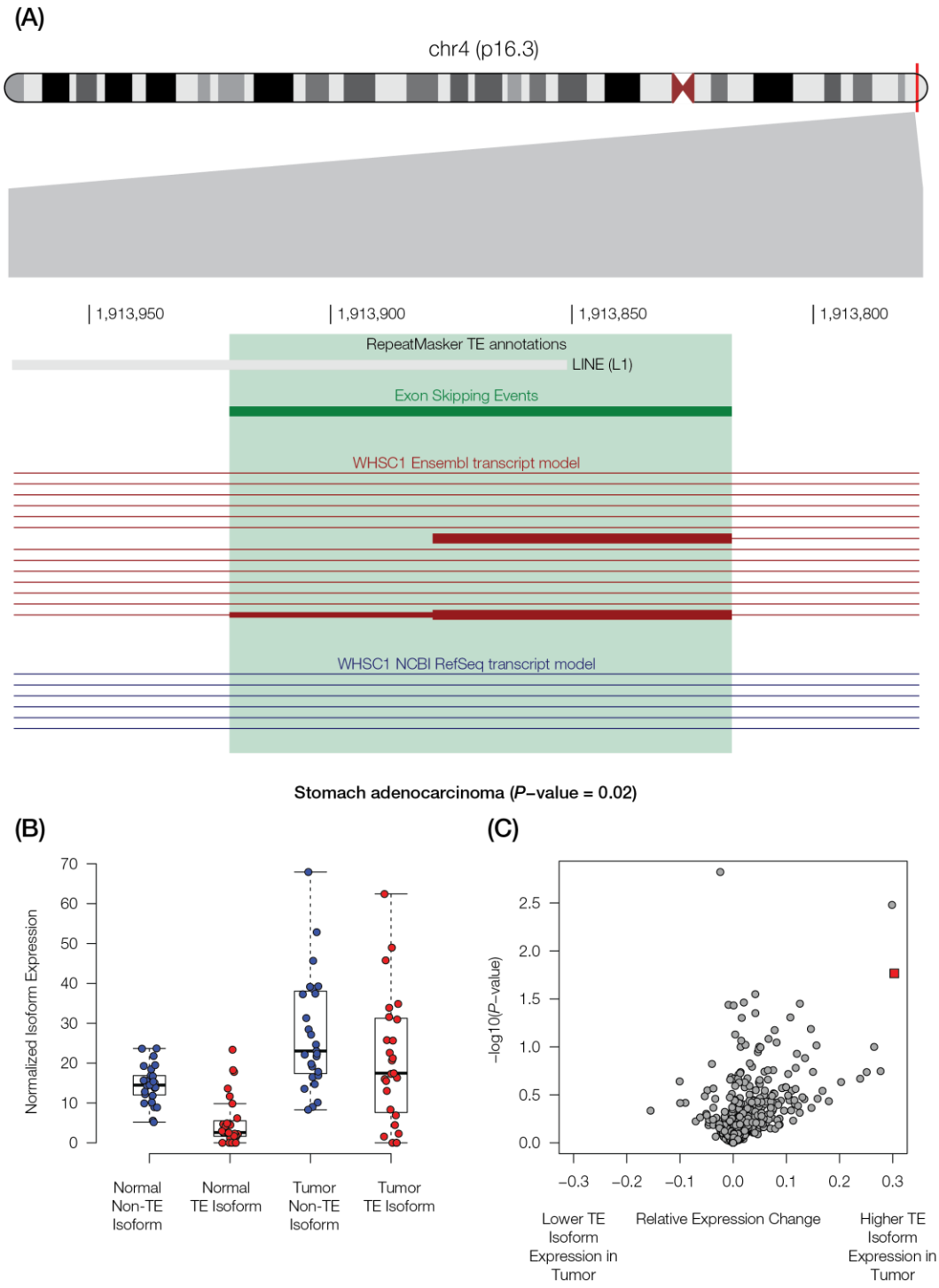

Figure 11 TE-derived alternative splicing in the *WHSC1* gene.

(A) The location of WHSC1 on the short arm of chromosome 4 is shown along with the specific location of its TE-derived alternative splicing event. A LINE (L1) sequence results in an exon skipping event. (B) Distributions of the non-TE (blue) and TE-derived (red)

isoforms are shown for matched normal (left) and stomach adenocarcinoma samples (right). (C) Relative expression change (REC) values are plotted against the corresponding G-test P-values (see Methods and Figure 30) for the matched normal and stomach adenocarcinoma samples. The WHSC1 TE-derived isoform values are shown as a red square.

The Calcium Activated Nucleotidase 1 encoding gene *CANT1* is overexpressed in prostate cancer and thought to be involved in proliferation, DNA synthesis, cell cycle, and migration of prostate cancer cells [143]. *CANT1* is known to undergo alternative splicing, with three well-defined isoforms. Here, we observe a novel exon skipping event, which appears to be induced by both SINE and LINE elements and results in a differentially expressed isoform, found at 32.7% in normal samples and 62.6% in stomach adenocarcinoma tumor samples (Figure 31).

3.5 Conclusions

Our global survey of TE-induced alternative splicing in cancer revealed that TE sequences contribute widely to alternate splice sites in cancer-associated genes, including cases where the TE isoforms are relatively overexpressed in tumor tissue. We hope that the landscape of TE-derived splice sites uncovered by our study can serve as a resource for further investigations into the role of TEs in tumorigenesis, and we have created a database of the TE-derived splice sites discovered here to facilitate follow-up studies on TE-induced alternative splicing. The data are distributed as a 'Track data hub' [144] on the UCSC Genome Browser at:

http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&hubUrl=http://jordan.biology.gatech.edu/teAs/hub.txt

The tracks show the genomic locations of the TE-induced alternative splicing events, with a separate track for each of the four splicing event types. The tracks can be used for visual inspection of individual events of interest or for more large-scale studies via download with the Table Browser.

One of the more intriguing results uncovered by our study was the potential connection between TE-induced alternative splicing and cancer fusion genes. Tumorigenesis is often characterized by large-scale genome rearrangements, and cancer fusion genes are thought to result from translocations, which bring genes that are normally far apart in the genome into close physical proximity. Our results showed numerous alternatively spliced exons that correspond to gene fusion junctions, particularly for the KLK2 gene that experiences both promiscuous alternative splicing and several gene fusion events, and these exons have previously been implicated in gene fusion events. We propose a model whereby apparent gene fusions actually occur at the transcript level via trans-splicing facilitated by TE sequences.

Pre-mRNA sequences destined for splicing are bound by heterogeneous ribonucleoprotein particle (hnRNP) proteins, which prevent the formation of short secondary structures caused by base pairing of complementary regions in the pre-mRNAs. In this way, the bound hnRNPs ensure that pre-mRNAs remain accessible for the assembly of the spliceosome. It occurs to us that hnRNP bound pre-mRNAs will also be open to trans interactions with pre-mRNAs from different loci, if they possess complementary sequences. Trans-splicing is the phenomenon whereby the splicing machinery joins splice donor and acceptor sites from different pre-mRNAs that are co-bound in the same spliceosome, yielding fused mature mRNAs. We propose that TE dispersed repeats

provide complementary sequences for binding between pre-mRNAs from different loci, thereby serving as hot spots for trans-splicing. We envision this mechanism as an RNA level analog of ectopic recombination between dispersed TE DNA sequences and a potential driver of transcriptome diversity.

It is important to note that our model of TE-induced trans-splicing for the generation of fusion transcripts is speculative and only suggested by our data. A number of additional analyses would need to be conducted to validate this model. DNA sequence analysis is needed to distinguish genome level rearrangements in cancer tissue from transcript fusions. TE homology (i.e. sequence complementarity) between transcript fusion partners, co-located with fusion junctions, would need to be confirmed. Explicit reconstruction of entire fusion transcript models, as opposed to individual alternative splice event analysis as was done here, needs to be performed to fully characterize observed gene fusions. Finally, it will be important to avoid RNA-seq experimental artifacts caused by template switching during the cDNA generation step, which could be done via single molecule RNA-sequencing.

CHAPTER 4. TUMOR SUPPRESSOR GENES AND ALLELE-SPECIFIC EXPRESSION: MECHANISMS AND SIGNIFICANCE

4.1 Abstract

LoF TSG alleles are shown to be segregating in world-wide populations of normal/healthy individuals at remarkably high frequencies, thereby establishing the potential importance of these genes in pre-disposing otherwise healthy individuals to cancer. To directly evaluate the possible contribution of the ASE of tumor suppressor LoF alleles in cancer onset/progression, matched sets of normal and tumor tissues isolated from 233 cancer patients representing 4 diverse tumor types were analyzed. The results indicate that there are functionally important changes in patterns of ASE in individuals heterozygous for LoF TSG alleles associated with cancer onset/progression. While a variety of molecular mechanisms were identified as potentially contributing to changes in ASE patterns in cancer, changes in DNA copy number and allele-specific alternative splicing mediated by anti-sense RNA emerged as a predominant factors.

4.2 Introduction

The long-standing belief that cancer is a genetic disease driven by mutations in a select set of oncogenes and/or tumor suppressor genes (aka, "cancer driver" genes) [145-149], has been augmented in recent years to incorporate the auxiliary contribution of changes in a variety of regulatory controls [150-152]. Recent findings indicate that these

additional regulatory controls may, in at least some instances, manifest as allele-specific expression (ASE) at specific cancer driver gene loci [153, 154]. ASE is the phenomenon whereby two or more gene alleles are differentially expressed with respect to one another [155, 156]. The potential clinical consequences of ASE have been previously documented [157, 158] including emerging evidence for the potential contribution of ASE to cancer [154, 159].

If cancer driver mutations can be transcriptionally repressed/de-repressed in an allele-specific manner, it follows that cancer driver mutations may be necessary but not always sufficient for onset and progression of the disease. For example, cancer driver mutations may, to a greater or lesser extent, be repressible and thus segregating at higher than expected frequencies in populations of normal healthy individuals. In addition, regulatory modulations in the ASE of cancer driver mutations may themselves, in at least some instances, be a significant contributor to cancer onset and progression. Of particular interest, in this regard, are those genes where loss-of-function (LoF) mutations have been shown to drive cancer onset/progression. This class of cancer driver genes is commonly known as tumor suppressor genes (TSGs) because a functional wild-type allele is considered sufficient to "suppress" the cancer driver effect of LoF alleles in heterozygotes. While LoF tumor suppressor mutations are typically considered to be recessive [160, 161], if these mutant alleles can be significantly differentially expressed relative to wild-type alleles in heterozygotes, the clinical consequences could be significant.

In this study, we first demonstrate that LoF TSG alleles are segregating in world-wide populations of normal/healthy individuals at remarkably high frequencies, thereby establishing the potential importance of these genes in pre-disposing otherwise healthy

individuals to cancer. To directly evaluate the possible contribution of ASE of tumor suppressor LoF alleles in cancer onset/progression, we analyzed matched sets of normal and tumor tissues isolated from 233 cancer patients representing four diverse tumor types. The results indicate that there are functionally important changes in ASE in individuals heterozygous for LoF TSG alleles associated with cancer onset/progression. While a variety of molecular mechanisms were identified as potentially contributing to changes in ASE in cancer, changes in DNA copy number and allele-specific alternative splicing mediated by anti-sense RNA emerged as a predominant factor.

4.3 Results

4.3.1 Tumor Suppressor Mutations are Abundant in Human Populations

The Catalogue Of Somatic Mutations In Cancer (COSMIC) is the world's largest database of somatic mutations associated with cancer onset and progression [72]. To determine the extent to which cancer associated mutations are segregating in the general human population, the genomic locations of all coding mutations in COSMIC census genes were intersected with sequence variants identified in individuals comprising the Phase 3 release of the One Thousand Genomes Project (1KGP) . The Phase 3 release catalogues all of the genetic variants present in 2504 putatively healthy individuals, representing a diversity of racial and ethnic groups randomly selected from 26 human populations around the world.

Remarkably, all individuals in the 1KGP were found to contain at least 31 homozygous and 68 heterozygous COSMIC census mutations (Figure 32). In total, 2,296 and 3,123 COSMIC census mutations were found in oncogenes and tumor suppressor genes, respectively, in healthy individuals. However, since the functional significance of all COSMIC mutations is not yet known and the fact that gain-of-function (dominant) mutations are difficult to unambiguously identify [162, 163], we focused our subsequent analyses on COSMIC mutations in TSGs that could be definitively classified as deleterious (i.e., non-sense, frame-shift, deletion mutations), along with all missense mutations predicted to be damaging by both The Sorting Intolerant from Tolerant (SIFT) [164] and Polymorphism Phenotyping v2 (PolyPhen-2) [165] algorithms. Employing this more conservative metric, 448 LoF COSMIC census mutations (Table 11) in TSGs were identified, of which ~93% of individuals carried at least one (Figure 12a). These 448 LoF mutations mapped to 137 different TSGs in at least one individual and 4 of these TSGs: Cbl Proto-Oncogene C (*CBLC*), Cadherin 11 (*CDH11*), Leucine Zipper Like Transcription Regulator 1 (*LZTR1*), and Tet Methylcytosine Dioxygenase 2 (*TET2*) had LoF mutations in >25% of the population (Figure 12b). Collectively, these findings indicate that genetic variants previously characterized as "cancer driver" mutations are segregating at relatively high frequencies in populations of individuals not afflicted with the disease.

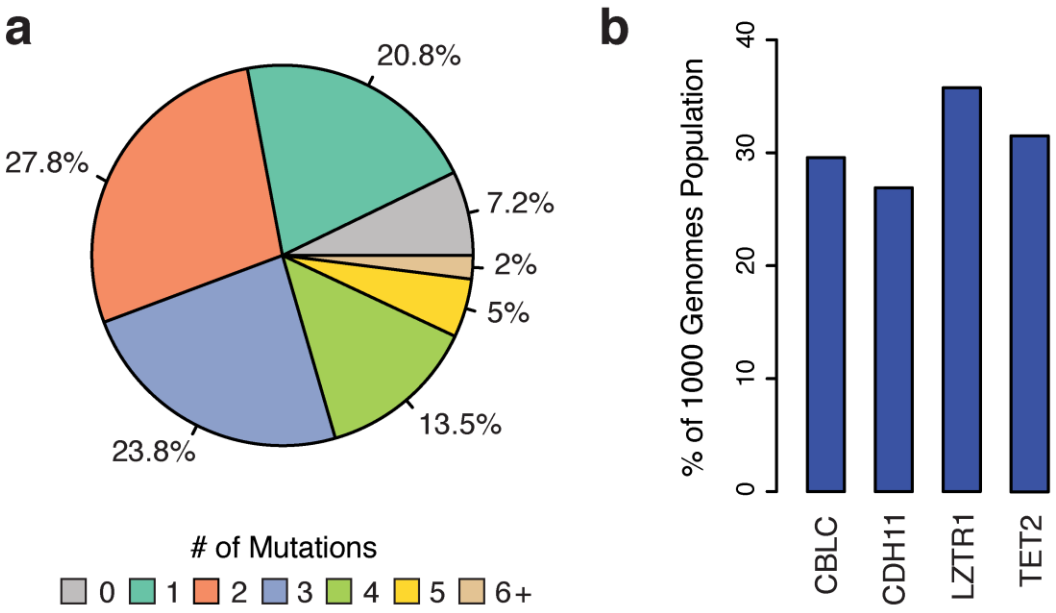

Figure 12 Distribution of LoF COSMIC census mutations in TSGs of the 1KGP.

Cancer associated mutations were identified in the 1000 genomes population (1KGP) as detailed in the Materials and Methods. a, Pie chart depicting the percent of the 1KGP containing deleterious cancer associated mutations in at least one TSG. b, Four TSGs most frequently mutated (LoF) in 1KGP.

4.3.2 A Minority (<20%) of TSGs Display Genetic Profiles in Cancer Consistent with Knudson's Two-Hit Hypothesis

Given the relative abundance of TSG LoF alleles in human populations, we utilized the The Cancer Genome Atlas (TCGA) database [166] to explore the possible contribution of these genes to cancer onset and/or progression by examining matched sets of cancer and normal tissues collected from 233 cancer patients representative of four diverse cancer types (breast invasive carcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma, and thyroid carcinoma). According to a model first proposed by Alfred Knudson in 1971 [167], newly arising LoF TSG mutant alleles, being recessive, can be

carried by normal cells with little significant negative effect. According to this model, acquisition of a second LoF mutation in the alternate wild-type allele is pre-requisite for tumor onset.

To test this hypothesis in our dataset, we genotyped all samples and identified TSGs that were heterozygous for a LoF mutation in normal tissues but that have acquired a secondary LoF mutation in the wild-type allele in the tumor samples. In total we found that only 46 of the 233 cancer patients (19.7%) were associated with acquisition of homozygosity in cancer for LoF alleles at TSG loci consistent with Knudson's "two-hit" hypothesis. These results indicate that the vast majority of TSGs heterozygous for wild-type and LoF alleles in normal tissues remain heterozygous in tumor tissue. However, if recessive LoF alleles can be significantly overexpressed relative to the wild-type alleles in an ASE fashion, LoF TSGs may be significant contributors to cancer onset/progression even in the heterozygous state.

4.3.3 The Proportion of LoF Mutations Displaying ASE is Significantly Elevated in Cancer Tissues

To explore the possible contribution of ASE in matched sets of normal and cancer tissues, we employed DNA-seq data from the TCGA database to identify all heterozygous sites in the exome and subsequently leveraged complementary RNA-seq data to compare the expression of wild-type or "reference" (ref) vs LoF mutant or "alternative" (alt) alleles at those loci (Figure 33).

The proportion of COSMIC census mutations (SNPs-single nucleotide polymorphisms) in TSGs displaying ASE was found to be significantly higher in the cancer relative to normal tissues for breast invasive carcinoma, head and neck squamous cell carcinoma, and lung adenocarcinoma ($P < 3.11 \times 10\text{-}10$) (Figure 13). Thyroid carcinoma was the only cancer type not displaying a significant difference, perhaps because these cancers are typically associated with a relatively low mutation rate [168]. To determine if this regulatory change was limited to TSG loci, we computed ASE for all heterozygous SNPs (het-SNPs) exome-wide. We found that all genes, on average, contain a significantly higher proportion of het-SNPs displaying ASE in breast, lung, head & neck ($P < 3.46 \times 10\text{-}15$) and thyroid ($P < 0.005$) tumors than normal samples (Figure 13). Thus, dysregulation in cancer, at least as manifest by ASE, is not limited to TSGs but extends to genes not previously identified as being implicated in tumorigenesis.

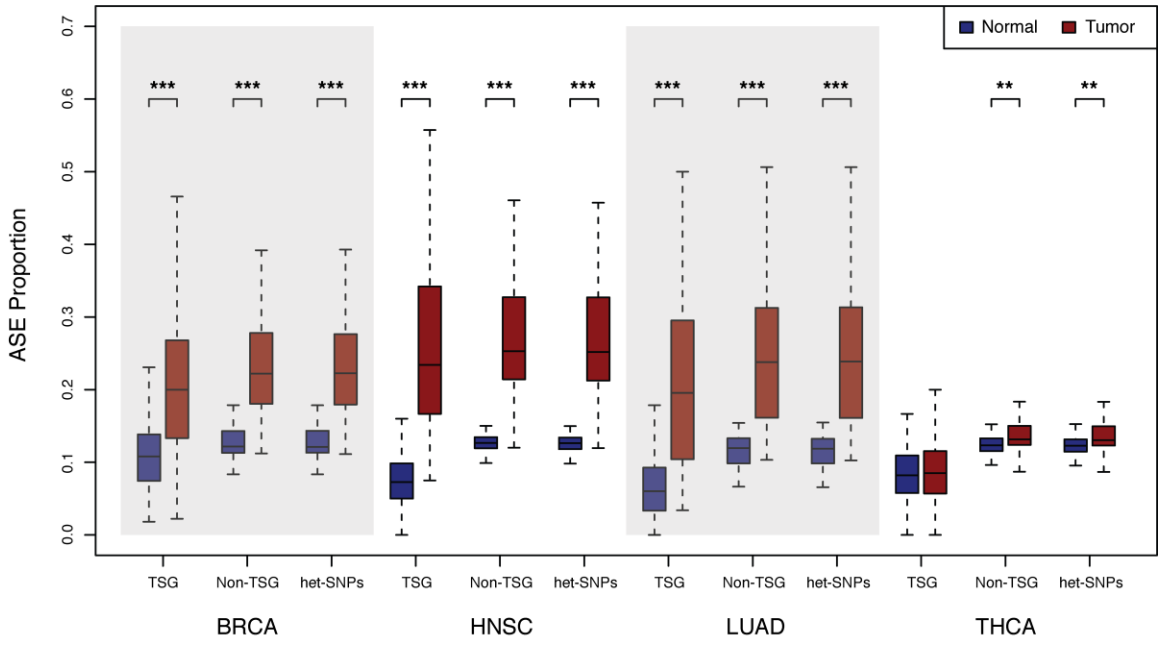

Figure 13 Distribution of the Proportion of ASE Loci.

Allele counts were generated for normal and primary tumor tissue pairs for breast invasive carcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma and thyroid carcinoma via analysis of RNA-Seq as described in the Materials and Methods section. Boxplots show the distribution of the of the proportion of COSMIC Census mutations in TSGs (left), all heterozygous SNPs in Non-TSGs (center) and all het-SNPs (right) with significant ASE (FDR = 5%, P < 0.005) in normal (blue) and tumor (red) samples.

4.3.4 Differences in Patterns of ASE Between Normal and Tumor Tissues Includes but is not Limited to TSGs

Changes in the relative expression of wild-type (ref) alleles vs. mutant (alt) alleles between normal and cancer tissues may manifest in one of six alternative ASE patterns: Pattern 1: No significant difference in ASE (ref=alt) in normal tissues but significant ASE (ref<alt) in cancer tissues; Pattern 2: Significant ASE in normal tissues (ref>alt) but no significant ASE (ref=alt) in cancer tissues; Pattern 3: Significant ASE in normal sample (ref>alt) and significant ASE in tumor sample (ref<alt); Pattern 4: No significant ASE in

normal tissues (ref=alt) but significant ASE in cancer tissues (ref>alt); Pattern 5: Significant ASE (ref<alt) in normal tissues but no significant ASE in cancer tissues (ref=alt); Pattern 6: Significant ASE in normal (ref<alt) and in cancer tissues (ref>alt) (Figure 14a). Patterns 1-3 are potentially of the most significance to cancer onset/progression because, in each case, the expression of the cancer driver LoF mutant (alt) allele is expressed at a significantly higher level than the wild-type allele in cancer tissues.

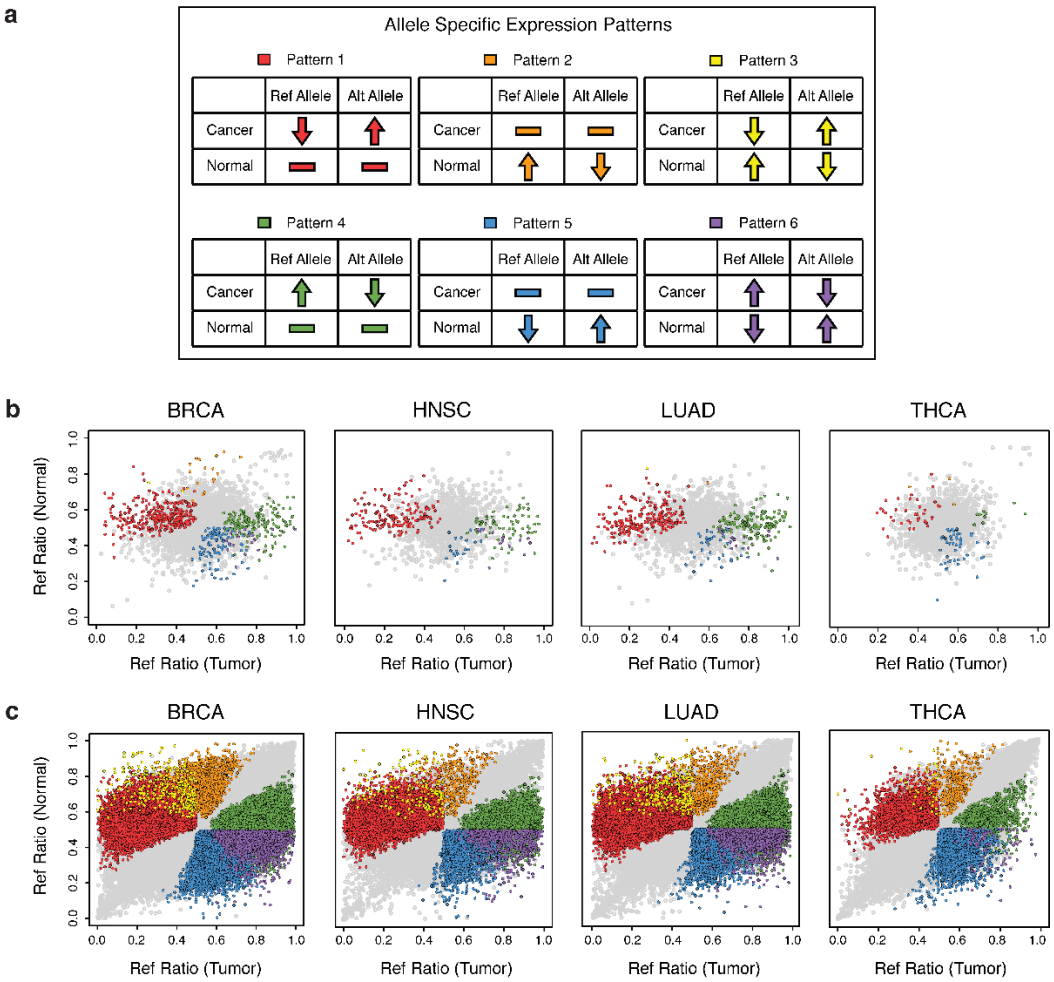

Figure 14 ASE SNP Patterns.

Allele counts were generated for normal and primary tumor tissue pairs for breast invasive carcinoma, head and neck squamous cell carcinoma, lung adenocarcinoma and thyroid carcinoma via analysis of RNA-Seq as described in the Materials and Methods section. Sites demonstrating significantly different ASE ratios (P < 0.05) between normal and tumor sample pairs that follow a tumorigenesis model are color coded by expression pattern as demonstrated in panel a. a, Six ASE patterns of interest were analyzed; Pattern 1: no significant ASE in normal sample and significant ASE (ref<alt) in tumor; Pattern 2: significant ASE (ref>alt) in normal sample and no significant ASE in tumor; Pattern 3: significant ASE in normal sample (ref>alt) and significant ASE in tumor sample (ref<alt); Patterns 4-6 mirror Patterns 1-3 with the opposite allele over expressed. Significant ASE (FDR = 5%, P < 0.005) was determined using a binomial test within samples in order to group loci into patterns. b, Reference allele ratios (ref/total) for all COSMIC Census loci in TSGs intersecting normal and tumor sample pairs, for 233 TCGA participants are shown here. c, Reference allele ratios (ref/total) for all loci intersecting normal and tumor sample pairs, for 233 TCGA participants are shown here.

The observed changes in ASE between matched sets of normal and cancer tissues for each of our 233 patients grouped into their respective Patterns, is presented in Table 5. A significant percentage of mutations in TSGs were found to display various patterns of ASE (FDR = 5%, $P < 0.005$; breast 14.9%, head and neck 16.0% and lung 19.6%) with Patterns 1 and 4 being the most predominant (see also Figure 14b). Thyroid cancer again stood out as an outlier where only 4.1% of mutations in TSGs were found to display ASE with Patterns 1 (2.3%) and 5 (1.3%) being nearly equally abundant.

Table 5 Percent of SNPs displaying ASE in 233 TCGA patients.

	Pattern	% Total SNPs	% All COSMIC	% TSG	% Oncogene	% Fusion
BRCA	1	7.3	7.6	7.3	7.0	8.3
	2	0.4	0.8	0.3	0.4	1.4
	3	0.1	0.2	0.0	0.0	0.3
	4	3.8	4.1	3.7	3.4	4.8
	5	2.6	2.5	2.8	2.2	2.7
	6	0.5	0.5	0.8	0.4	0.4
	No ASE	85.2	84.4	85.1	86.6	82.1
HNSC	1	8.7	9.7	10.5	10.1	8.3
	2	0.4	0.4	0.0	0.0	0.9
	3	0.2	0.3	0.0	0.0	0.6
	4	4.6	4.5	3.7	4.2	4.7
	5	2.1	1.7	1.3	1.2	2.0
	6	0.7	1.0	0.5	0.9	1.0
	No ASE	83.4	82.5	83.9	83.5	82.4
LUAD	1	9.7	10.7	9.2	9.8	12.0
	2	0.3	1.0	0.1	0.0	2.0
	3	0.2	1.0	0.0	0.0	1.9
	4	5.4	6.8	7.0	6.5	6.9
	5	1.9	2.3	2.2	1.6	2.7
	6	0.7	0.8	1.1	0.6	0.9
	No ASE	81.7	77.4	80.4	81.5	73.5
THCA	1	2.1	2.3	2.3	1.7	2.6
	2	0.2	0.5	0.0	0.1	1.0
	3	0.0	0.0	0.0	0.0	0.0
	4	0.5	0.8	0.5	0.4	1.1
	5	1.6	1.9	1.3	2.0	2.1
	6	0.1	0.0	0.0	0.0	0.0
	No ASE	95.5	94.5	95.9	95.8	93.2

When the analysis was extended to include all transcribed genes ("%Total SNPs" in Table 5), a similar trend was observed, where 14.8%, 16.6% and 18.3% of all het-SNPs were found to display ASE in breast, head and neck and lung cancers, respectively. Thyroid cancer was again an outlier with only 4.5% of all transcribed genes displaying ASE (Figure 14c). Collectively these results indicate that changes in ASE in cancer are widespread and not limited to TSGs.

To explore this apparent dysregulation of COSMIC census mutations in TSGs further, we aggregated our SNP ASE data to quantify ASE of the entire allele of a gene by employing the Meta-analysis Based Allele-Specific Expression Detection (MBASED) protocol [154]. We found 14.4%, 17.9%, 20.4% and 5.7% of all TSGs show ASE in breast, head and neck, lung and thyroid cancers, respectively (Table 22). These results are consistent with the relative levels of ASE detected within SNPs in these cancers with Pattern 1 again emerging as a predominant pattern (9.1%, 11.1%, 13.2%, and 2.8%) (Table 22).

One example of those TSGs displaying changes in patterns of ASE in cancer is the Human Leukocyte Antigen A1 gene (HLA-A). HLA-A has been previously identified as a hotspot for ASE activity [169], likely due to the high genetic variability that is well-documented in the major histocompatibility complex [170, 171]. We detected changes in ASE in the HLA-A gene in 20% of our patients including nucleotide positions not previously reported to display ASE [172].

Another example is Tumor Protein P53 (TP53) that displayed the highest level of ASE in our breast cancer patients (57.9% of all patients) displaying Pattern 4 63.6% of the

time (Figure 15). Additionally, breast cancer implicated TSGs Breast cancer type 1 susceptibility protein (*BRCA1*) and Cadherin 1 (CDH1) were found to display changes in ASE in 15.4% and 32.4% of breast tumors, respectively, frequently displaying Pattern 1 (Figure 15). Interestingly, Zinc Finger Protein 331 (ZNF331) was the only TSG predominately displaying Pattern 2 (Figure 15). A previous study [173] has shown ZNF331 to display large amounts of ASE in breast cancer, citing genomic imprinting as a possible explanation [174].

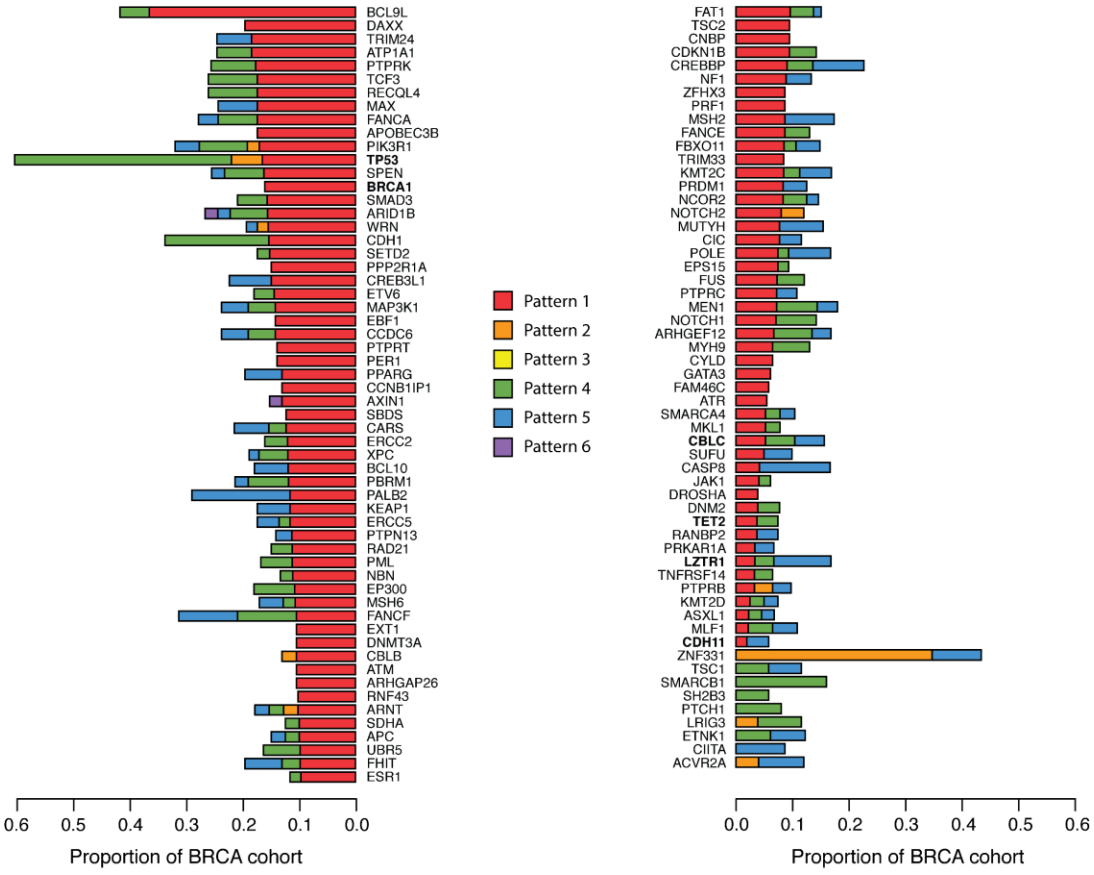

Figure 15 Tumor suppressor genes with ASE in breast cancer patients.

Gene level ASE was computed as described in the Materials and Methods section. The proportion of breast cancer patients with ASE in 115 TSGs.

The four TSGs: *CBLC*, *CDH11*, *LZTR1*, and *TET2* previously shown to be most frequently mutated in 1KGP (Figure 12b) were also observed to display changes in ASE in breast cancer (Figure 15). Similar trends in the frequency of ASE Patterns among TSGs were observed in head and neck, lung, and thyroid cancers, with thyroid again sporting the least amount of ASE (Figure 34).

4.3.5 Changes in DNA Allelic Ratios May Explain up to 35% of the Observed Changes in ASE Between Normal and Cancer

Perhaps the most straight-forward explanation of the observed changes in ASE in cancer is that it is a reflection of underlying changes in allele counts on the DNA level. For example, it is known that the duplication or deletion of alleles on the DNA level can contribute to ASE in cancer [153, 175]. In addition, the polyclonal heterogeneity of most tumors can manifest as an imbalance in DNA allele counts and associated ASE changes in analyses carried out on bulk tumor samples.

To explore the extent to which changes in DNA copy number may be contributing to the observed ASE in our samples, we downloaded whole-exome sequencing data (WXS) for nine randomly selected patients representing each of the three cancer types displaying the highest level of ASE (breast invasive carcinoma, lung adenocarcinoma, and head & neck squamous cell carcinoma) (Table 23). We ensured these individuals displayed ASE in COSMIC genes (Figure 35) and that their ASE was evenly spread throughout the genome (Figure 36). We found that, on average, 35.2% (45.7% Breast, 25.8% Head and Neck, and 20.5% Lung) of ASE genes displayed DNA allele counts that correlated with

RNA allele counts (Table 6). Further, investigation of these samples revealed that only 10% of these genes displayed copy number duplications potentially accounting for their ASE (Figure 16a). Collectively these findings indicate that while, on average, a large fraction of our observed changes in ASE may be accounted for by corresponding changes in DNA allele counts, a significant fraction of ASE in cancer is likely attributable to allele-specific changes in gene regulation.

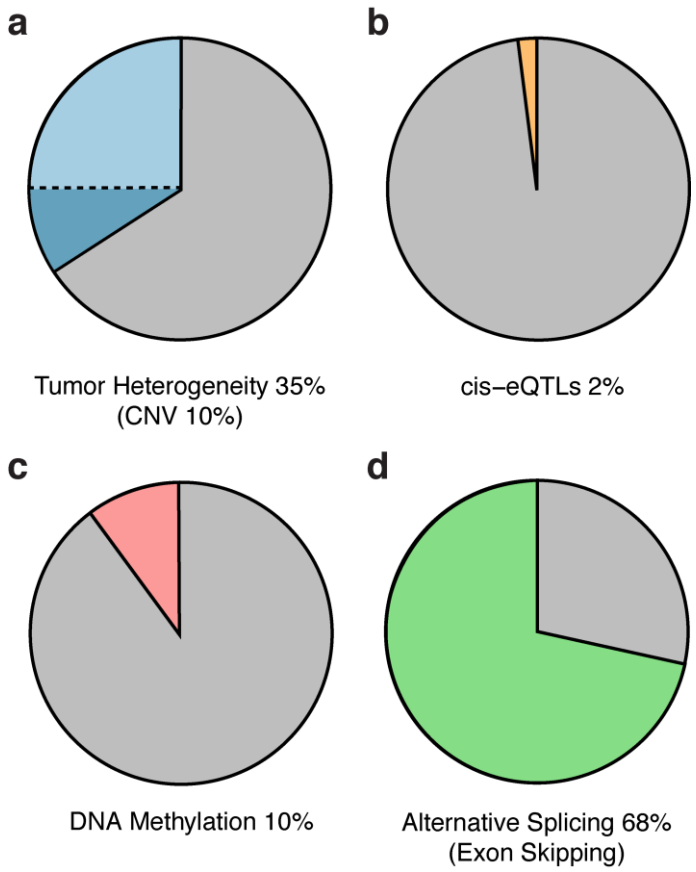

Figure 16 Mechanisms of ASE.

Potential underlying mechanisms for ASE were explored as outlined in the Materials and Methods. Pie chart depicts amount of ASE which could be attributed to a, tumor heterogeneity and copy number variation (CNV) b, cis-eQTLs, c, DNA methylation and d, exon-skipping via computational analysis.

Table 6 ASE patterns potentially explained by DNA counts.

Patient	ASE SNPs explained by DNA counts	Total ASE SNPs	Percentage of ASE correlated
Breast 1	526	1336	39.4
Breast 2	719	1411	51.0
Breast 3	177	367	48.2
Head & Neck 1	275	930	29.6
Head & Neck 2	179	674	26.6
Head & Neck 3	217	993	21.9
Lung 1	56	233	24.0
Lung 2	47	155	30.3
Lung 3	11	167	6.6
Total	**2207**	**6266**	**35.2**

4.3.6 Allele-specific cis-Regulatory Variation may Account for a Small Fraction of Observed Changes in ASE Between Normal and Cancer

Allele-specific regulatory changes in gene expression could be explained by sequence variation mapping to cis-regulatory regions located up- or down-stream of affected genes [20, 176, 177]. To explore the extent to which allele-specific cis-regulatory variation may account for ASE in cancer, we identified expression-quantitative trait loci (eQTLs) present in six of our nine patients' normal and tumor samples using the Genotype-Tissue Expression Project's (GTEx) single tissue cis-eQTL data available for breast and lung tissue [178]. eQTLs are regions of the genome containing DNA sequence variants previously established to regulate gene expression levels [179]. Genes previously established to be regulated by at least one eQTL are classified as eGenes [180, 181].

Genes displaying ASE in our study were found to be significantly enriched for eGenes relative to genes not displaying ASE ($P = 0.018$) (Figure 37). This finding was pronounced for breast ($P = 2.56 \times 10\text{-}6$) and lung cancer ($P = 6.22 \times 10\text{-}4$) patients (Figure

37). However, collectively only 24% of genes displaying ASE in our dataset are eGenes and just 3% of ASE eQTLs are ASE-specific (i.e., found only in genes o ASE). Moreover, we found that the expression slope of an eQTL is not often correlated with the allelic expression of a gene (1.8%; Figure 16b; Table 25). For example, consider the heterozygous eQTL variant (rs10654) mapping to the 3' UTR of the NUP54 gene in both normal and tumor samples of breast cancer patient 2 (TCGA-BH-A0BW). Despite being heterozygous, this eQTL is not associated with ASE in the normal tissue but is associated with ASE in cancer tissue where the alternative haplotype is overexpressed and in phase with the highly expressed alternative eQTL allele (Figure 38a).

We also pursued the eQTLs differing in genotype between normal and tumor samples for specific evidence of cis-regulation. While infrequent, we did find several notable cases where eQTL genotypes correlated with ASE. Shown in Figure 38b, is a model for how cis-eQTLs may be responsible for the intragenic ASE we observed. In this particular example, three separate eQTLs within a 50bp region (rs34176173, rs12085114, rs34016668) are found ~4.8k base pairs from the 3’ UTR of the gene NME7 in breast invasive carcinoma patient 3 (TCGA-BH-A0DT). The eQTL is homozygous for the ref allele in the normal sample that does not show ASE, and heterozygous in the tumor sample. The eQTL alternative allele that is associated with high expression of NME7 is present on the alt haplotype being overexpressed. Further, all three eQTLs are in linkage disequilibrium with the ASE SNP ($r > 0.42$) suggesting they segregate together. We found four additional cases where cis-eQTLs could account for ASE but none of these were associated with COSMIC census genes.

Collectively, the above findings indicate that while allele-specific cis-regulatory variation may account for some instances of ASE, it alone does not explain the vast majority (>75%) of instances of ASE in our dataset.

4.3.7 Changes in Methylation may Account for a Small Fraction of Changes in ASE Between Normal and Cancer

Another possibility is that ASE is regulated epigenetically. For example, it has been previously suggested that epigenetic inactivation of one of the two alleles could result in ASE [182]. Epigenetic effects across chromosomes are often regionally associated with CpG repeats or "CpG islands" [183-185]. To determine if genes displaying ASE in our dataset display evidence of regional chromosomal clustering, we visualized the genomic locations of ASE for nine patients on a genome ideogram (Figure 36). The results provide no evidence for regional chromosomal clustering indicative of regional epigenetic effects.

To further search for evidence of epigenetic involvement in ASE in our dataset, we analyzed global DNA methylation sites in normal and cancer tissues since this is a common mechanism by which gene transcription can be repressed epigenetically [186-188]. Methylation data were downloaded from TCGA for seven of the nine patients described above and used to compare genes that had a significant change in methylation with genes showing a significant change in ASE. We found that only 10.2% of genes displaying ASE also displayed significant differences (>1.3-fold) in methylation between normal and tumor tissues (Figure 16c; Table 25). Although these results indicate that changes in methylation are not likely to be playing a significant role in the ASE detected in our patient samples,

the analysis cannot be considered definitive because the methylation data provided by TCGA are not allele specific.

4.3.8 A Significant Fraction of Changes in ASE Between Normal and Cancer may be a Reflection of Underlying Alternative Splicing Events Induced by Anti-sense RNA

A recent study has implicated allele-specific alternative splicing as a potentially significant factor in ASE [189]. For example, consider a scenario where an allele-specific exon-skipping event occurs more frequently in a cancer tissue than normal (Figure 39). This would result in a negligible difference in the level of transcripts containing the wild-type (ref) and LoF mutant (alt) allele in normal but significantly fewer transcripts containing the wild-type allele ("T allele" in Figure 39) in cancer.

To explore the possibility that allele-specific alternative splicing may be contributing to the observed ASE in our patient samples, we leverage previously computed isoform counts for TCGA patient data [122]. Specifically, we sought to determine if there is a significant increase in exon skipping in genes displaying ASE. The results indicate that 70% of SNPs displaying changes in ASE between normal and cancer correlate with an increased frequency of exon-skipping events (i.e., >1.5-fold increase in expression of reads consistent with exon-skipping events) (Table 3; Figure 16d).

While these results suggest that allele-specific alternative splicing may be a significant contributor to ASE, it does not provide a mechanism as to how two variant alleles from the same gene may be alternatively spliced. One possibility is that the point

mutations or indels that distinguish mutant LoF (alt) alleles from wild-type (ref) alleles map to consensus splice sites or other cis-regulatory locations known to be involved in the splicing process [111, 121]. However, of the 100,852 SNPs associated with changes in ASE between normal and tumor, only 1.4% (1,418/100,852) map to consensus splice sites (716 in acceptor G, 702 in donor AG).

A second possible mechanism that may explain how two variant alleles from the same gene may be alternatively spliced emerges from previous studies showing that splicing events can be experimentally induced in vivo by exposing primary transcripts to even small fragments of anti-sense RNAs that pair with known splice sites in the primary transcript [14, 190]. We reasoned that if such allele-specific anti-sense RNAs are being differentially produced in normal and cancer tissues, it may explain observed differences in allele-specific splicing and consequent differences in ASE.

To test this hypothesis, we quantified the levels of anti-sense RNAs mapping to splice sites adjacent to allele-specific alternative-splice events. The results presented in Table 3 demonstrate a significant increase in levels of anti-sense RNA in genes displaying allele-specific alternative-splice events associated with ASE. For example, Figure 17a depicts a case where the *ADAM15* gene displays ASE in breast cancer patient 3 (TCGA-BH-A0DT). The ADAM15 protein is known to display tumor suppressive activities when it is released as an exosomal component [191], and abnormal expression and dysregulation of alternative splicing in *ADAM15* has been previously associated with breast cancer [192-194]. Previous studies have also shown that four ADAM15 isoforms varying by the sequence of the cytoplasmic domains, display variable effects in vitro. The shortest isoform, ADAM-15D, arises due to loss of exons 19 to 21 causing a reading frame shift in

exons 22 and 23 when compared with the other three isoforms. The variant lacks proline-rich modules and has a distinct sequence of 37 amino acids. As shown in Figure 17b, we observe an increase in anti-sense RNA mapping to acceptor (1.7x) and donor (1.8x) sites in this patient's tumor. We have identified an exon skipping event (exon 19), consistent with the ADAM-15D isoform. The increase in anti-sense RNA correlates with this isoform's expression, which is substantially higher (3.8x) in the patient's tumor sample when compared to normal and could explain ASE at this locus (Figure 21c).

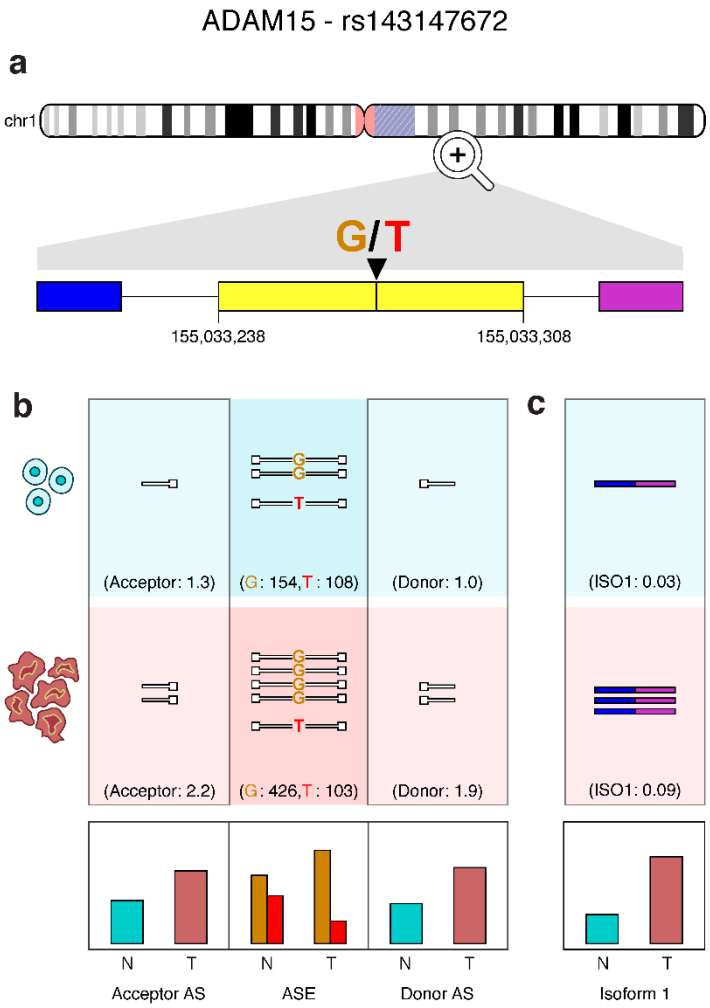

Figure 17 *ADAM15* exon skipping correlates with ASE in a breast adenocarcinoma patient.

a, An exon skipping event in exon 19 of ADAM15 in a breast cancer patient (TCGA-BH-A0DT). b, Antisense reads mapping to donor and acceptor sites are quantified, alongside the ASE locus within the exon. c, Quantification of reads supporting the isoform missing exon 19. Relative expression plots are shown for antisense RNA, ASE and isoforms below.

Another example is illustrated in Figure 40a, where the Lysyl Oxidase Like 2 (*LOXL2*) gene displays ASE at the rs1051146 locus in breast cancer patient 1 (TCGA-BH-A0B3) which overlaps with an exon skipping event. *LOXL2* has accumulated numerous reports that document its role in cancer formation and proliferation of breast cancer [195-197]. Further, research has shown that a short isoform of LOXL2 missing exon 13 can

regulate cancer cell migration and invasion through a dissimilar mechanism compared to its canonical form [198]. Here, we observe more anti-sense RNA mapping to acceptor (8.7x) and donor (9.2x) sites (Figure 40b) and increased skipping of exon 6 (9.9x) (Figure 40c) in breast cancer patient 1's tumor sample, both correlated with an increase in ASE.

Tenascin C (*TNC*) is a gene belonging to a family of extracellular matrix (ECM) glycoproteins that is known to be overexpressed in cancer cells. Studies have shown that remodeling of ECM in cancer can affect cellular interaction as ECM influences behavior of the cells [199, 200]. One specific study has shown that a TNC isoform containing exons 14 and 16 but not 15 is upregulated in breast cancer which leads to increased cell invasion and proliferation [201]. In breast cancer patient 3, *TNC* displays changes in ASE at rs17819466 inside exon 15 (Figure 41a). Anti-sense RNA mapping to acceptor (2.1x) and donor (2.6x) sites are elevated in the tumor sample (Figure 41b), as are split-reads spanning exons 14 and 16 (5.5x) (Figure 41c).

Collectively, these results suggest that anti-sense mediated alternative splicing is a significant factor in accounting for our observed changes in ASE between normal and cancer and may be contributing to tumorigenesis.

4.4 Discussion

Cancer is a complex disease not only from the perspective of the number and diversity of genes involved but also because of the existence of extensive regulatory variation controlling the expression of these genes. One manifestation of these regulatory

controls is allele-specific expression (ASE) at specific cancer driver gene loci [153]. If cancer driver mutations can be transcriptionally repressed/de-repressed in an allele-specific manner, they may be segregating at higher than expected frequencies in populations of normal healthy individuals. In an initial effort to explore this possibility, we conducted a computational analysis of functionally significant cancer driver mutations in a sampling of normal healthy human populations across the world (2.5 thousand genomes comprising the 1000 Genome Project (1KGP) [59]). While relatively few confirmed dominant oncogene mutations were found to be segregating in these populations, a remarkable 93% of healthy individuals sampled were found to carry functionally significant loss-of-function (LoF) cancer driver mutations at one or more tumor suppressor gene loci (21% of individuals carry 1 mutant allele; 28% carry 2, 24% carry 3, 13% carry 4, 5% carry 5, 2% carry >6). While these frequencies are higher than what have been generally reported for specific TSGs [202, 203], they are not unprecedented. Among the most intensively studied TSGs is the RB1 gene that is associated with inherited childhood retinoblastoma [167]. Although the frequency of individuals heterozygous for LoF RB1 alleles ("carriers") in human populations is generally reported to be □ 5% [203], considerable variability exists among ethnic groups/populations. For example, in a study of select Asian populations, the frequency of carriers of LoF RB1 alleles was reported to be as high as 34% in specific ethnic populations [204]. Collectively, our findings serve to confirm the potential importance of TSGs in pre-disposing a significant fraction of otherwise healthy individuals to cancer.

The fact that LoF TSG alleles are typically recessive to their "partner" functional or "wild-type" allele lead to the "two-hit" hypothesis first proposed by Alfred Knudson in

1971 [167]. The "two-hit" hypothesis proposes that individuals heterozygous for a LoF tumor suppressor allele will not typically develop cancer unless an additional LoF mutation occurs in the gene's functional partner allele. While the two-hit hypothesis has been successfully employed to account for many instances of inherited cancers associated with tumor suppressor genes [205-207], a number of examples have been identified in recent years that are inconsistent with Knudson's "two-hit" hypothesis [208-211]. For example, it is now known that not all children afflicted with retinoblastoma are homozygous for the LoF RB1 allele [212, 213] and this condition has, in several cases, been associated with aberrant expression of unlinked regulatory genes [214].

To evaluate the "two-hit" model in our dataset, we computed the frequency of patients heterozygous for a LoF (alt) mutation in normal tissues that acquired a secondary LoF mutation in the partner wild-type (ref) allele in matched sets of cancer and normal samples. We found that < 20 % of patients acquired a second LoF mutation in cancer tissues as predicted by the "two-hit" model. This finding is consistent with a growing body of evidence that the mechanisms underlying the contribution of TSGs to cancer onset and progression are often more complex than originally envisioned [215, 216].

A primary goal of our current study was an evaluation of the potential significance of changes in ASE of TSGs between normal and cancer, and to explore the molecular mechanisms that may underly this process. Towards this end, we searched the TCGA database for evidence of ASE in matched sets of normal and tumor tissue samples isolated from 233 randomly selected cancer patients representing four diverse tumor types. We found that COSMIC census mutations in TSGs display significantly ($P < 3.11 \times 10\text{-}10$) more ASE in tumors compared to matched normal tissues in breast, head and neck, and

lung cancers. Our finding that this change is not limited to COSMIC genes but extends to genes not previously associated with cancer, implies a general loss of regulatory control in cancer. Evidence for such a global loss in regulatory control in cancer has been previously reported [217-219]. Although the mechanisms underlying these global changes in regulatory control in cancer remain largely unknown, epigenetic changes are emerging as one potentially important player in the process [220].

In a preliminary effort to explore the possible contribution of epigenetics to the global changes in ASE observed between our normal and cancer samples, we downloaded methylation data for our patient samples from TCGA. We found that only 10% of genes displaying ASE also displayed significant differences (>1.3-fold) in methylation between our normal and tumor tissue samples. While changes in methylation are generally considered to be a reliable indicator of epigenetic-associated changes in gene expression [221], our results suggest that changes in methylation are not likely to be playing a significant role in the regulation of ASE in our patient samples.

Of the six possible Patterns of change in ASE between normal and cancer tissues, we found that Pattern 1 (i.e., no ASE in normal tissue but expression of mutant allele (alt) > expression of wild-type (ref) allele in cancer tissue) was one of the most commonly observed Patterns across cancer types. This finding is consistent with the hypothesis that LoF TSG alleles may be contributing significantly to cancer onset/progression even in the heterozygous state.

One possible explanation of the observed changes in ASE between normal and cancer tissue is that it is structural in nature, i.e., the consequence of differences in allele

counts attributable to, for example, loss of heterozygosity (LOH) or the polyclonal heterogeneity characteristic of most tumors [175]. To test this possibility, we compared RNA allele counts with DNA allele counts in the same patient samples. We found that on average 35% of genes displaying ASE had DNA allele counts that correlated with RNA allele counts. These results are consistent with prior findings indicating that a significant fraction of ASE can be accounted for by underlying differences in DNA allelic content [175]. However, only 10% of these genes displayed copy number duplications consistent with their ASE. Thus, at least with respect to our patient samples, differences in ASE between normal and cancer tissues is not merely structurally based but likely attributable to allele-specific differences in gene expression.

Allele-specific differences in gene expression may be attributable to variant cis-regulatory sequences located up- or down-stream from the respective alleles' coding regions. Such cis-regulatory variation is commonplace and is often identified by utilizing QTL mapping methodologies [222]. We employed the Genotype-Tissue Expression Project's (GTEx) single tissue cis-eQTL database to explore the extent to which allele-specific cis-regulatory variation may account for ASE in our patient samples. We found that only 24% of genes displaying ASE are eGenes, more of which explain ASE in normal (38%) than tumor (21.8%) samples. Moreover, only 1.8% of ASE haplotypes were found to be in phase with an eQTL indicating that cis-regulatory variation is not a likely explanation of the majority of instances of ASE in our dataset.

Having failed to identify a mechanism of transcriptional level regulation that could explain the majority of observed instances of ASE in our dataset, we turned our attention to the potential influence of post-transcriptional regulation on ASE. One post-

transcriptional mechanism of growing prominence in cancer biology is alternative splicing (AS) [12, 223]. The primary RNA products of genes are processed at the post-transcriptional level by alternative RNA splicing resulting in multiple RNA isoforms per gene. If alternate RNA isoforms are generated on an allele-specific basis (allele-specific alternative splicing), it could manifest itself as differences in ASE. To explore the possibility that allele-specific alternative splicing could be contributing to changing patterns of ASE in cancer, we examined isoform counts associated with our TCGA patient data [122]. We found that almost half (46%) of SNPs displaying ASE in our patient samples were indeed associated with exon skipping.

While the potential functional significance of alternative splicing in cancer has been long noted [123], the mechanisms underlying the phenomenon remain poorly understood. Because the genes displaying changes in ASE are not associated with cis-regulatory mutations in splice acceptor/donor sites, we focused our attention on possible trans-regulatory mechanisms. One possibility is that one or more of the regulatory proteins or RNAs associated with the spliceosome could be mutated or otherwise dis-regulated in cancer resulting in aberrant splicing patterns [120]. However, the fact that our observed allele-specific alternative splicing was limited to only a subset of genes suggested that the underlying mechanism was of a more targeted nature.

One possibility was suggested from previous studies showing that splicing events can be experimentally induced in vivo by exposing primary transcripts to anti-sense RNAs that pair with known splice sites in the primary transcript [16, 17, 224]. Indeed, there is growing evidence that de novo expression of anti-sense RNAs may play a significant role

in the induction of alternate splice variants [225] and that this may be a significant factor in cancer onset/progression.

4.5 Summary and Conclusions

We have shown that LoF TSGs are segregating in human populations at significant frequencies suggesting that many otherwise healthy individuals are at elevated risk of developing cancer. Changes in ASE between normal and cancer tissues indicates that LoF TSG alleles may contribute to cancer onset/progression even when heterozygous with wild-type functional alleles. While a variety of molecular mechanisms were identified as potentially contributing to changes in ASE between normal and cancer, differences in DNA counts and allele-specific alternative splicing mediated by anti-sense RNA emerged as predominant factors.

4.6 Materials and Methods

4.6.1 Cancer Associated Mutation Identification in 1000 Genomes Population

Using the BEDTools program [226], the genomic locations of all coding mutations in COSMIC census genes (v82) [130] were intersected with a vcf file containing all sequence variants called from the 2504 individuals of the Phase 3 release of the 1000 Genomes Project (1KGP) [59]. The distribution of these cancer associated mutations was determined for all intersecting mutations including the subset of deleterious mutations.

Variant effects were annotated using Variant Effect Predictor (VEP) using the Ensembl 91 release [227]. Mutations were considered to be deleterious if they were non-sense, frameshift, splice acceptor/donor mutations, or whole gene deletion mutations. Missense mutations predicted deleterious by both SIFT [164] and Polyphen2 [165] were also scored as deleterious mutations. Moreover, we removed any mutation that had been labeled as benign or likely benign by clinvar [228].

4.6.2 Sequencing Data Acquisition

Whole exome sequencing (WXS) and transcriptome sequencing (RNA-Seq) data for matched sets of normal and primary tumor samples collected from 233 cancer patients were downloaded from The Cancer Genome Atlas (TCGA) via Genomic Data Commons (GDC) (89 Breast Cancer (BRCA) patients, 52 Lung Adenocarcinoma (LUAD) patients, 39 Head and Neck Squamous Cell Carcinoma (HNCSC) patients and 28 Thyroid Cancer (THCA) patients; Table 26). The data consisted of paired-end reads generated from Illumina platforms. As per the harmonization pipeline of GDC all WXS samples were aligned to GRCh38 reference genome [229], indels were locally realigned and base quality scores were recalibrated. RNA-Seq samples were also aligned to GRCh38 by the GDC. Autosomes were subsequently extracted for downstream analysis using SAMtools [63]. To investigate the mechanisms of ASE Whole Genome Sequencing data (WGS) for nine of these patients (three breast invasive carcinoma, three head and neck squamous cell carcinoma and three lung adenocarcinoma) were downloaded. The BAM files for WGS

were checked for quality using FASTQC, lifted over to GRCh38 and local realignment of indels was performed using GATK [69].

4.6.3 Genotyping and Variant Calling with WXS and Variant Annotation

Genotyping was implemented from WXS. SAMtools mpileup output was fed to VarScan's mpileup2snp function in order to call variants [230]. Only reads with mapping quality > 14 were counted. Further, to call a variant, a position must have met a minimum read depth of 8, minimum allelic depth of 2 and variant allele frequency threshold of 0.2. The default p-value of 0.01 was used for calling variants. Variants were annotated using VEP with the same criteria mentioned as above.

4.6.4 Allele Specific Expression Analysis

4.6.4.1 Counting Allele-Specific Reads

Indexed RNA-Seq BAM files along with filtered heterozygous variants were passed to GATK's ASEReadCounter tool [69]. At this step, only reads with minimum mapping quality and base quality scores of 20 and 30, respectively, were counted. Also, minimum depths of 20 reads per site and four reads per allele were applied. With the aim of inferring biological significance, resulting allele counts were annotated with rsid using Kaviar. Subsequently, gene names associated with particular SNPs were fetched from dbSNP using EDirect [231]. The fraction of reads containing the reference allele over the

total number of reads at a given position (Ref Ratio) was calculated for all heterozygous SNPs. Custom scripts were written to perform allele specific expression analysis.

4.6.4.2 Accounting for Mapping Bias

When mapping RNA-seq reads to the reference genome, reads overlapping a SNP that contain the alternative allele tend to map less frequently than those reads containing the reference allele. This allelic mapping bias has been well documented and presents challenges in ASE analysis [232]. Degner et al. demonstrated that the reliability in ASE estimation is greatly dependent on the capability to control for reference mapping bias [233]. To limit this bias, we first removed sites known to be susceptible to mapping bias. We did so by removing all sites with 50bp mapability < 1 based on the UCSC mapability track [234]. To correct for any residual bias, we calculated the genome-wide allelic ratios for all nucleotide pairs and used them in place of 0.5 as the expected allelic ratio in the binomial test (Figure 42) as previously done by Lappalainen et al [235].

4.6.4.3 ASE Analysis

Using the allele counts for every heterozygous position that met our filtering requirements, we performed a binomial test to identify whether the ratio of reference and alternative read counts differed significantly from the corresponding expected proportion between those alleles. Expected ratios were inflated slightly from 0.5 based on the observed allele counts within our population as described in the previous section. We classified a

site as an ASE SNP if its binomial p-value was less than 0.005 and corrected for a false discovery rate (FDR) of 5%. Gene level ASE was determined by aggregating ASE information from all heterozygous SNPs within a gene as outlined by the MBASED protocol [154]; ASE genes were classified with a major allele frequency (MAF) greater than 0.7 and p-value less than 0.05 (FDR 5%). To label significant ASE genes with Patterns we pseudo-phased them by creating a major haplotype consisting of the alleles with higher RNA read counts. If a haplotype contained more reference SNPs it was labelled as the reference haplotype and vice versa for the alternative. If the number of reference and alternative SNPs on each haplotype were the same, the haplotype was labelled as ambiguous.

Differences in ASE between normal and cancer tissue groups, were evaluated by comparing the distributions of the proportion of SNPs with ASE within each collection. The statistical significance levels of the observed difference in ASE between normal and tumor tissues for both COSMIC census mutations and all heterozygous SNPs were evaluated by comparing these distributions using the non-parametric Mann-Whitney U test.

When comparing SNPs intersecting paired normal and tumor samples, we applied a combined binomial-Fisher test to determine if ASE patterns were significant. Three ASE patterns of interest were analyzed; Pattern 1: No significant difference in ASE (ref=alt) in normal tissues but significant ASE (ref<alt) in cancer tissues; Pattern 2: Significant ASE in normal tissues (ref>alt) but no significant ASE (ref=alt) in cancer tissues; Pattern 3: Significant ASE in normal sample (ref>alt) and significant ASE in tumor sample (ref<alt); Pattern 4: No significant ASE in normal tissues (ref=alt) but significant ASE in cancer

tissues (ref>alt); Pattern 5: Significant ASE (ref<alt) in normal tissues but no significant ASE in cancer tissues (ref=alt); Pattern 6: Significant ASE in normal (ref<alt) and in cancer tissues (ref>alt). All Patterns are visualized in Figure 14a. Significant ASE (FDR = 5%, $P < 0.005$) was determined using a binomial test within samples and subsequently a Fisher's exact test ($P < 0.05$) when comparing two samples. Both tests were applied to increase stringency and validity of results.

4.6.5 Second Site Loss-of-Function Mutations

Filtered heterozygous sites in tumor suppressor genes (TSGs) of all 233 patients in normal and tumor samples were phased using SHAPEIT [236]. Loss-of-function mutations were defined as stop gained, frameshift, splice acceptor/donor, start lost and stop lost mutations. Deleterious missense mutations predicted to be damaging/deleterious by SIFT [164] and Polyphen2 [165] were also considered loss of function in TSGs. Patients with a secondary site loss-of-function mutation were defined as having a heterozygous mutation in the normal sample and either: 1) the same mutation homozygous in the tumor sample, 2) a new loss of function mutation on the opposite allele in the tumor sample (i.e. compound heterozygote), or 3) a DNA segment with loss of allele at the locus in the tumor sample. Segments of DNA with loss of allele were identified using FACETS [237].

4.6.6 Analyses to Determine Mechanisms of ASE

4.6.6.1 Cis Expression Quantitative Trait Loci (cis-eQTL) Analysis

In order to investigate the possible contribution of upstream/downstream regulatory variation to ASE, cis-eQTL detection was performed. Variants were called, using SAMtools mpileup and bcftools [63], on the genomic sequence ± 1mb from the transcription start site (TSS) of each gene. Mono-allelic sites were phased using SHAPEIT [236] with the 1KGP reference panel of haplotypes for each participant's super population. Indels and multi-allelic sites were phased using HapCUT [238]. Subsequently, the outputs of the two tools were merged to complete the phasing step. Correlated cis-eQTL-gene pairs, for lung and breast tissues, were downloaded from The Genotype Tissue Expression Project's (GTEx) single tissue cis-eQTL data [181]. The eQTL-gene pairs were intersected with variants of the breast and lung cancer TCGA participant samples. We then created a list of all possible eQTL SNP – ASE SNP pairs, by pairing eQTL variants with all ASE SNPs in the corresponding eGene. We tested to see if there was a difference in the enrichment of eQTLs in ASE genes versus non-ASE genes using the Fisher's exact test, across all individual samples. Resulting eQTL SNP – ASE SNP pairs were tested for linkage disequilibrium (LD) using PLINK [239] and the appropriate super population from 1KGP (EUR or AFR) based on the patients' reported race. Pairwise correlations (r) for all SNP pairs were computed. Finally, VEP [227] was used to find eQTLs in regulatory regions and genomic regions of interesting cases were visualized using the AllelicImbalance R package [240].

4.6.6.2 Tumor Heterogeneity and Allele Specific Copy Number Variation

To explore the possible contribution of DNA copy number variation to ASE, DNA read counts and allelic depths were generated for heterozygous sites in matched normal and tumor samples using the snp-pileup utility provided in the FACETS package [237]. A binomial test with an expected ratio of 0.5 and an FDR of 5% were used to select sites where the allelic depth of the reference and alternative allele were significantly different. These sites were overlapped with ASE SNPs to see how much of the observed ASE could be attributed to the heterogenous nature of tumor tissues and copy number variations.

Segments with non-diploid copy number variations or evidence of loss of allele were identified using FACETS with a c value of 100. To calculate the amount of ASE associated with copy number variation, ASE SNPs were overlapped with non-diploid copy number variation segments to see if the copy number change supported the allelic imbalance observed in the RNA-Seq reads.

4.6.6.3 Methylation

To investigate the possible contribution of methylation to ASE, Illumina Infinium HumanMethylation450 (HM450) and Human Methylation 27 (HM27) Array data were downloaded from TCGA for six patients (TCGA-50-5932, TCGA-BH-A0B3, TCGA-BH-A0BW, TCGA-BH-A0DT, TCGA-CV-6959, TCGA-CV-7255) for tumor and normal samples. Methylation intensity was quantified by a beta-value calculated as the ratio of the methylated probe intensity and the sum of the methylated and unmethylated probes. Fold

changes of the beta-values were calculated between the tumor and normal samples; a fold change of > 1.33 was considered significant. To estimate how many changes in ASE between tumor and normal samples could be accounted for by epigenetic changes in methylation we overlapped genes that had a significant change in methylation with genes showing a change in ASE.

4.6.6.4 Alternative Splicing

To explore the possible contribution of post-transcriptional alternative splicing to ASE, a compressed alternative splicing dataset from a study [122] containing over 8,000 patients across 32 cancer types on TCGA was downloaded from GDC. Exon-skipping event data for 233 patients in this ASE study were matched and pulled from the dataset for further analyses.

In order to compare the expression of a genelet that provides evidence for an exon-skipping event that could be contributing to observed ASE in matched normal and tumor samples, the genomic coordinates of ASE SNPs were first intersected with the start and stop positions of all confirmed exon-skipping events using bedtools. Kahles et al. [122] define ISO1 and ISO2 genelets as isoforms with shorter and longer lengths, respectively. For exon-skipping events, the ISO1 genelet refers to the boundary connecting the exons adjacent to the one being skipped. For the filtered list of exon-skipping events that intersect with an ASE SNP, the difference in the number of multi-exon spanning reads (ISO1) between normal and its paired tumor sample were inspected for each of the ASE patterns. The isoform counts were normalized to counts per million (CPM) using the total number

of reads to account for sequencing depth. For ASE SNPs with Patterns 1 and 4, a SNP was counted as correlated with exon-skipping if a 1.5x fold increase in ISO1 CPM was observed from the tumor sample to the matched normal sample, *i.e.* ($\frac{tumor\ ISO1\ CPM}{normal\ ISO1\ CPM} \geq 1.5$). Similarly, for ASE SNPs with Patterns 2 and 5, a SNP was counted as correlated with exon-skipping if a 1.5x fold increase in ISO1 CPM was observed from normal to tumor, *i.e.* ($\frac{normal\ ISO1\ CPM}{tumor\ ISO1\ CPM} \geq 1.5$). For ASE SNPs with Patterns 3 and 6, evidence of exon-skipping was necessary in both samples, *i.e.* ($normal\ ISO1\ CPM > 0$, $tumor\ ISO1\ CPM > 0$).

4.6.7 Splice Site Mutations

Exonic regions of splice site motifs were first defined prior to examining ASE SNPs that could be contributing to the observed alternative splicing. These regions were defined as two bases downstream of the acceptor site (AG) and three bases upstream (5'-ward) of the donor site (GT). Strand information was taken into account while defining these exonic regions for genes on the minus strand. Using bedtools, ASE SNPs were intersected with the specified exonic regions.

4.6.8 Antisense RNA

Detecting antisense RNA requires the alignment to be performed with reads that have strand information available. The sequence alignment map (BAM) files available on

TCGA are missing such information as unstranded library kits were used to generate the reads. The CAFE [241] pipeline predicts the orientation of the unstranded reads using Markov chain models coupled with maximum likelihood estimation given a reference BAM file generated from strand-specific RNA-Seq reads. For the nine patients (3 LUAD: TCGA-44-6776, TCGA-50-5932, TCGA-55-6984; 3 BRCA: TCGA-BH-A0B3, TCGA-BH-A0BW, TCGA-BH-A0DT; 3 HNSC: TCGA-CV-7255, TCGA-CV-7416, TCGA-CV-6959), BAMs consisting of reads with predicted directions were generated using the pipeline and three cell lines with stranded RNA-Seq reads available for each cancer type (LUAD: HCC78 - SRR2050924; BRCA: MCF7 - SRR5048141; HNSC: neuroblastoma-derived cell line – SRR4787038).

Regions where antisense RNA could interfere with the splicing of an exon were determined using the canonical splicing motif for Homo sapiens [242] along with a gene's strand. (+ strand: acceptor= -40bp AG +2bp, donor= -3bp GT +5bp; - strand: acceptor= -2bp AG +40bp, donor= -5bp GT +3bp). Using bedtools, the reads that intersected these regions on the opposite strand of the coding gene were quantified in order to estimate the amount of antisense reads. The change in the number of antisense reads mapping to the splicing motif of genes between normal and tumor samples was quantified as a fold change in antisense expression. The read counts were normalized with CPM, using the total number of reads to account for sequencing depth.

CHAPTER 5. LEVERAGING TCGA GENE EXPRESSION DATA TO BUILD PREDICTIVE MODELS FOR CANCER DRUG RESPONSE

5.1 Abstract

5.1.1 Background

Machine learning has been utilized to predict cancer drug response from multi-omics data generated from sensitivities of cancer cell lines to different therapeutic compounds. Here, we build machine learning models using gene expression data from patients' primary tumor tissues to predict whether a patient will respond positively or negatively to two chemotherapeutics: 5-Fluorouracil and Gemcitabine.

5.1.2 Results

We focused on 5-Fluorouracil and Gemcitabine because based on our exclusion criteria, they provide the largest numbers of patients within TCGA. Normalized gene expression data were clustered and used as the input features for the study. We used matching clinical trial data to ascertain the response of these patients via multiple classification methods. Multiple clustering and classification methods were compared for prediction accuracy of drug response. Clara and random forest were found to be the best clustering and classification methods, respectively. The results show our models predict with up to 86% accuracy; despite the study's limitation of sample size. We also found the

genes most informative for predicting drug response were enriched in well-known cancer signaling pathways and highlighted their potential significance in chemotherapy prognosis.

5.1.3 Conclusions

Primary tumor gene expression is a good predictor of cancer drug response. Investment in larger datasets containing both patient gene expression and drug response is needed to support future work of machine learning models. Ultimately, such predictive models may aid oncologists with making critical treatment decisions.

5.2 Background

The goal of personalized medicine is to tailor treatments for individuals based on unique characteristics of their genetic background. Given the vast variety of cancers and the inherent molecular heterogeneity of the disease, personalized medicine in cancer can be particularly effective, [243]. By studying molecular profiles of tumors, one can potentially discover biomarkers for drug sensitivity, resistance, or adverse effects that may be helpful in predicting drug response [244, 245]. Recent successes demonstrated small molecule inhibitors which target pathways upregulated in cancer patients [246]. Further, breast cancer has long served as a model for successful personalized oncology, by ¬administering treatments specific for HER2-positive patients [247].

While personalized oncology has shown signs of promise, not all cancers have such well-defined targetable pathways [23]. This has led to the recent emergence of machine learning for predicting cancer drug response. This method, though promising, has had rather limited success. Difficulties creating reliable predictive models have stemmed from a lack of clinical data to use for training, poorly annotated drug responses, and noise introduced by a large number of features [26]. In previous studies, the lack of patient data was offset by utilizing genomic and transcriptomic profiles of cancer cell lines as the features for predicting response to chemotherapy [25, 248-251]. Low interpretability and limited accuracy are the drawbacks of predicting in vivo response based on in vitro data. The high dimensionality of molecular data is prone to overfitting and can lead to deceptive associations from intrinsically multiplex gene networks. Together, these challenges have muddled attempts to build informative and accurate patient drug predictive models.

To address these complications, we applied several machine learning techniques. First, to reduce dimensionality we utilized optCluster [252], an R package for determining the optimal clustering algorithm and optimal number of clusters. OptCluster identifies highly similar or repetitive expression patterns from genes and clusters them into gene modules. This method reduces the number of features while also minimizing the amount of information loss. Secondly, we predicted drug response using the random forest algorithm [253] in order to protect against overfitting; a common issue with many machine learning methods. Random forest is an ensemble method which builds decision trees. This approach deters overfitting by incorporating a variety of features and leveraging a majority vote when performing classification [230]. Perhaps, most importantly, we evade using cell-line data to extrapolate in vivo predictions by instead harnessing gene expression data

available from primary tumor tissues in The Cancer Genome Atlas (TCGA) [57]. We observe robust prediction using our model and we evaluate predictive gene modules that are implicated in biological pathways critical to drug response.

5.3 Results

5.3.1 Drug Selection Results

Our study is based on data obtained from the National Cancer Institute's TCGA database [166]. TCGA provided both patients' clinical trial data and transcriptomic data from patients' primary tumor samples. This data included expression levels for 60,483 genes including protein-coding genes, non-coding RNA genes and pseudogenes. We used the Genomic Data Commons API to download the Upper-Quartile Normalized Fragments per Kilobase per Million mapped reads (UQ-FPKMs) from the patients' primary tumor samples. The clinical trial data consisted of 12,051 records with data for 32 cancer types. Each record contains clinical trial data for an individual patient. There are multiple clinical trials in the database and a patient will have one record for each clinical trial of which they participated (i.e. a patient can have two records if they participated in different trials). Each record included information about: drugs administered, patient demographics, temporal data of the study, and response of the patient.

For the purposes of classification, we defined a responder as a patient who had partial or complete response and a non-responder as a patient who had a clinical progressive disease or stable disease response. One pan-cancer model for each drug was created by including all the cancer types that the selected drug had treated. Only patients

on single drug therapy throughout the entire duration of treatment were retained in the study.

Based on these criteria, Fluorouracil (5-FU) and Gemcitabine (GCB) were chosen because they provided the highest number of records. Our study included two models: (1) 5-FU pan-cancer and (2) GCB pan-cancer. See Table 7 for the counts of each model.

Table 7 Patient counts for each model by response.

Model	Responder Count	Non-responder Count	Total Count
Fluorouracil pan-cancer*	34	24	58
Gemcitabine pan-cancer**	37	55	92

*Fluorouracil pan-cancer includes: colon adenocarcinoma, esophageal carcinoma, pancreatic adenocarcinoma, rectum adenocarcinoma, stomach adenocarcinoma

**Gemcitabine pan-cancer includes: bladder urothelial carcinoma, breast invasive carcinoma, cervical squamous cell carcinoma and endocervical adenocarcinoma, cholangiocarcinoma, head and neck squamous cell carcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, ovarian serous cystadenocarcinoma, pancreatic adenocarcinoma, pheochromocytoma and paraganglioma, sarcoma, skin cutaneous melanoma, testicular germ cell tumors, uterine corpus endometrial carcinoma

5.3.2 OptCluster Results

We report the results of the most accurate clustering algorithm from optCluster in Table 8. Clara coupled with several classification algorithms provided the best gene modules with a cross-validation mean accuracy of 84.1% (sd:10.7%) for 5-FU and 82.3% (sd: 8.6%) for GCB. In Figure 18, we see how accuracy changes relative to the number of selected clusters. The peak accuracy was with 32 and 50 clusters for 5-FU and GCB, respectively. We tested other classification methods, support vector machines and logistic

regression, but they yielded worse results. Cross-validation accuracy for support vector machine was 81% for 5-FU and 71.5% for GCB and logistic regression was 77.0% for 5-FU and 73.0% for GCB. There was minimal impact on accuracy when including demographic data of the patients (gender, age, cancer type and cancer stage) [5-FU: 83.6%; Gemcitabine: 79.1%]. Additional tuning improved the model validation accuracy, as seen in

Table 8.

Table 8 Number of clusters and mean accuracy for pan-cancer models.

	Fluorouracil			Gemcitabine		
Clustering Method	**# Considered Clusters****	**# Selected Clusters*****	**Average Accuracy**	**# Considered Clusters****	**# Selected Clusters*****	**Average Accuracy**
Optcluster (Clara) with RF*	204	32	84.1%	192	50	82.3%
Optcluster (Clara) with SVM*	204	32	81.0%	192	50	71.5%
Optcluster (Clara) with logistic regression	204	32	77.0%	192	50	73.0%
Optcluster (Clara) with RF and demographics	204	32	83.6%	192	50	79.1%
Model Validation (Clara)	204	32	52.9%	192	50	82.1%
Model Validation (Clara with Tuning)	204	32	52.9%	192	50	85.7%

* RF is for random forest; SVM is for support vector machine

**Number of considered clusters represents the number of clusters entered into random forest for variable importance ranking.

***Number of selected clusters is number of clusters selected by random forest for

classification

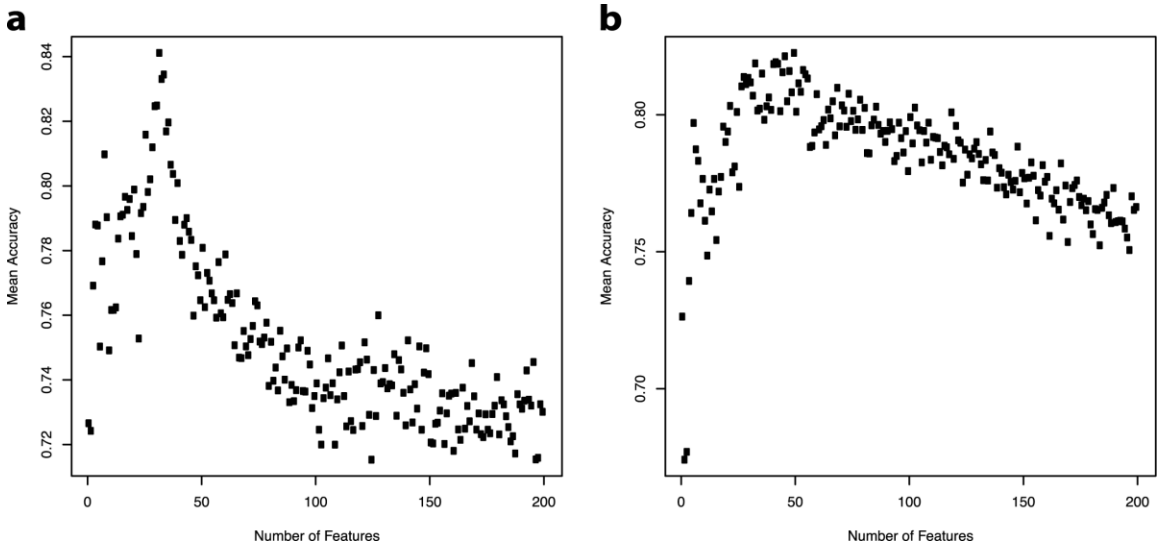

Figure 18 Accuracy of random forest by number of clusters (using clara clustering algorithm).

(A) Mean accuracy (200x cross-validation) was calculated using 1-204 clusters in order of importance from the 5-FU pan-cancer model. (B) Mean accuracy (200x cross-validation) was calculated using 1-192 clusters in order of importance from the GCB pan-cancer model.

In Figure 19a-b, the cross-validation predicted probabilities for non-responders and responders are plotted. The 5-FU model is particularly strong at identifying responders (97% sensitivity) while the GCB model is better at classifying non-responders (100% specificity).

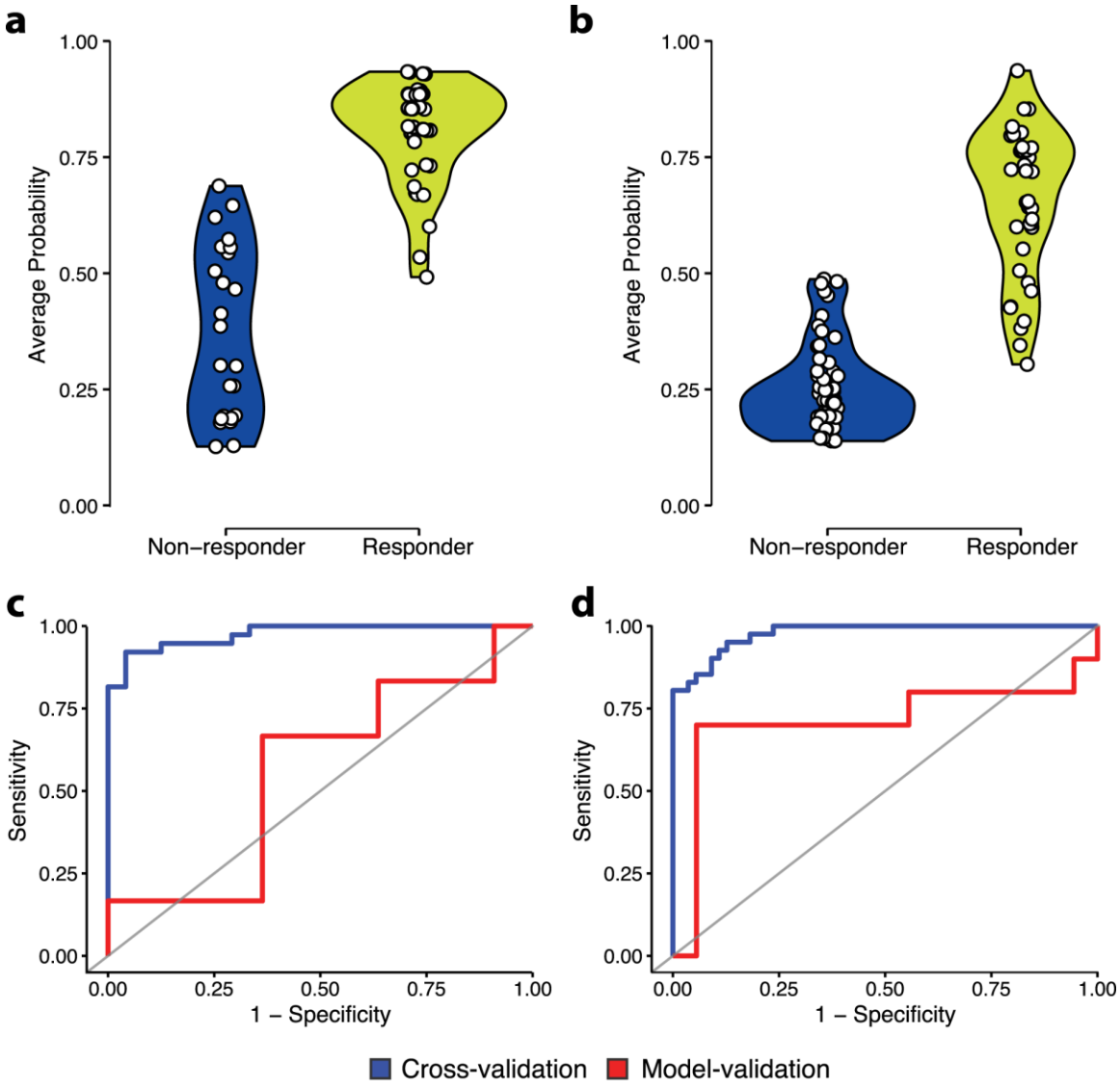

Figure 19 Random forest classifier performance for pan-cancer models.

(A) Comparison of the cross-validation predicted probabilities between non-responders and responders for the 5-FU pan-cancer model. (B) Comparison of the cross-validation predicted probabilities between non-responders and responders for the GCB pan-cancer model. (C) ROC curve for 5-FU pan-cancer model; Cross-validation (Sensitivity: 0.97 Specificity: 0.66 AUC: 0.98) Model-validation (Sensitivity: 0.64 Specificity: 0.33 AUC: 0.56). (D) ROC curve for GCB pan-cancer model; Cross-validation (Sensitivity: 0.80 Specificity: 1.0 AUC: 0.98) Model-validation (Sensitivity: 0.70 Specificity: 0.94 AUC: 0.71).

5.3.3 Model Validation & ROC Curves

The ROC curves in Figure 19c-d demonstrate the results of the sensitivity and specificity for the cross-validated accuracy of the training data and the model validation of the test data. The results of the model validation showed an increase in the accuracy for GCB by 3 percentage points to 85.7%. We did not see the same improvement in 5-FU, which dropped to 52.9% accuracy. The training data for both models performed with AUC = 0.98. More interestingly, we see the GCB validation curve still classifies well with AUC = 0.71, while the 5-FU validation curve is barely above the random classifier line; showing it doesn't perform much better than chance. This decrease in accuracy could be attributed to sample size and difficulty in predicting on the cancer type (stomach adenocarcinoma) used for this validation.

5.3.4 Gene Set Enrichment Analysis

Once we established a ranking of modules in terms of predictive importance, we performed a gene set enrichment analysis using PANTHER [254]. We analysed the genes that comprised the clara gene modules that the optimal models used for prediction (Table 27). The top 20 biological pathways by gene percent are shown for the pan-cancer models in

Table 9. In addition, Figure 20 shows the relationship of gene expression level between responders and non-responders for the each of these pathways. A high positive number, such as in P00018 indicates that the mean gene expression for the responders in

pathway, P00018, are higher than that of the non-responders. The opposite holds true for the negative values.

Table 9 Top 20 PANTHER pathways in models by gene percent.

Model	*Category name*	*Accession*	*# Genes*	*Percent of gene hit against total # genes*	*Percent of gene hit against total # pathway hit*	*raw P Value*	*FDR*
5-Flourouracil	Wnt signaling pathway	P00057	10	3.20%	11.80%	2.82E-02	4.60E+00
	Integrin signaling pathway	P00034	6	1.90%	7.10%	6.60E-02	3.59E+00
	Gonadotropin-releasing hormone receptor pathway	P06664	5	1.60%	5.90%	4.06E-01	2.88E+00
	Inflammation mediated by chemokine and cytokine signaling pathway	P00031	4	1.30%	4.70%	7.91E-01	3.39E+00
	Huntington disease	P00029	4	1.30%	4.70%	1.70E-01	2.78E+00
	p53 pathway	P00059	3	1.00%	3.50%	1.49E-01	2.69E+00
	EGF receptor signaling pathway	P00018	3	1.00%	3.50%	4.66E-01	3.16E+00
	PDGF signaling pathway	P00047	3	1.00%	3.50%	4.88E-01	3.18E+00
	Cadherin signaling pathway	P00012	3	1.00%	3.50%	5.08E-01	3.18E+00
	CCKR signaling map	P06959	3	1.00%	3.50%	7.45E-01	3.28E+00
	Apoptosis signaling pathway	P00006	2	0.60%	2.40%	6.97E-01	3.34E+00
	Angiogenesis	P00005	2	0.60%	2.40%	1.00E+00	1.00E+00
	Ionotropic glutamate receptor pathway	P00037	2	0.60%	2.40%	1.74E-01	2.36E+00
	FGF signaling pathway	P00021	2	0.60%	2.40%	7.03E-01	3.27E+00
	B cell activation	P00010	2	0.60%	2.40%	2.74E-01	2.80E+00
	DPP signaling pathway	P06213	1	0.30%	1.20%	1.10E-01	2.56E+00
	DPP-SCW signaling pathway	P06212	1	0.30%	1.20%	9.70E-02	2.64E+00
	BMP/activin signaling pathway-drosophila	P06211	1	0.30%	1.20%	1.10E-01	2.24E+00
	Axon guidance mediated by Slit/Robo	P00008	1	0.30%	1.20%	1.10E-01	2.24E+00
	Metabotropic glutamate receptor group III pathway	P00039	1	0.30%	1.20%	1.10E-01	2.24E+00
G e	Inflammation mediated by chemokine and cytokine	P00031	33	2.80%	6.70%	2.28E-05	9.27E-04

signaling pathway						
Wnt signaling pathway	P00057	28	2.40%	5.70%	1.85E-02	3.02E-01
T cell activation	P00053	21	1.80%	4.30%	2.95E-07	2.40E-05

Table 9 (continued).

Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway	*P00027*	*17*	*1.50%*	*3.50%*	*1.74E-03*	*5.68E-02*
B cell activation	P00010	17	1.50%	3.50%	1.82E-06	9.91E-05
Integrin signaling pathway	P00034	16	1.40%	3.30%	1.12E-01	8.33E-01
Angiogenesis	P00005	15	1.30%	3.10%	9.70E-02	7.91E-01
Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway	P00026	15	1.30%	3.10%	6.25E-02	7.27E-01
Cadherin signaling pathway	P00012	15	1.30%	3.10%	5.56E-02	6.97E-01
Nicotinic acetylcholine receptor signaling pathway	P00044	14	1.20%	2.90%	2.66E-03	7.22E-02
Gonadotropin-releasing hormone receptor pathway	P06664	13	1.10%	2.70%	1.00E+00	1.01E+00
EGF receptor signaling pathway	P00018	13	1.10%	2.70%	9.59E-02	8.22E-01
Apoptosis signaling pathway	P00006	10	0.90%	2.00%	2.29E-01	1.24E+00
CCKR signaling map	P06959	11	0.90%	2.20%	6.17E-01	1.60E+00
Parkinson disease	P00049	10	0.90%	2.00%	8.77E-02	8.41E-01
Interleukin signaling pathway	P00036	11	0.90%	2.20%	1.98E-02	2.93E-01
PDGF signaling pathway	P00047	9	0.80%	1.80%	7.21E-01	1.53E+00
FGF signaling pathway	P00021	9	0.80%	1.80%	4.27E-01	1.42E+00
Huntington disease	P00029	8	0.70%	1.60%	1.00E+00	1.00E+00

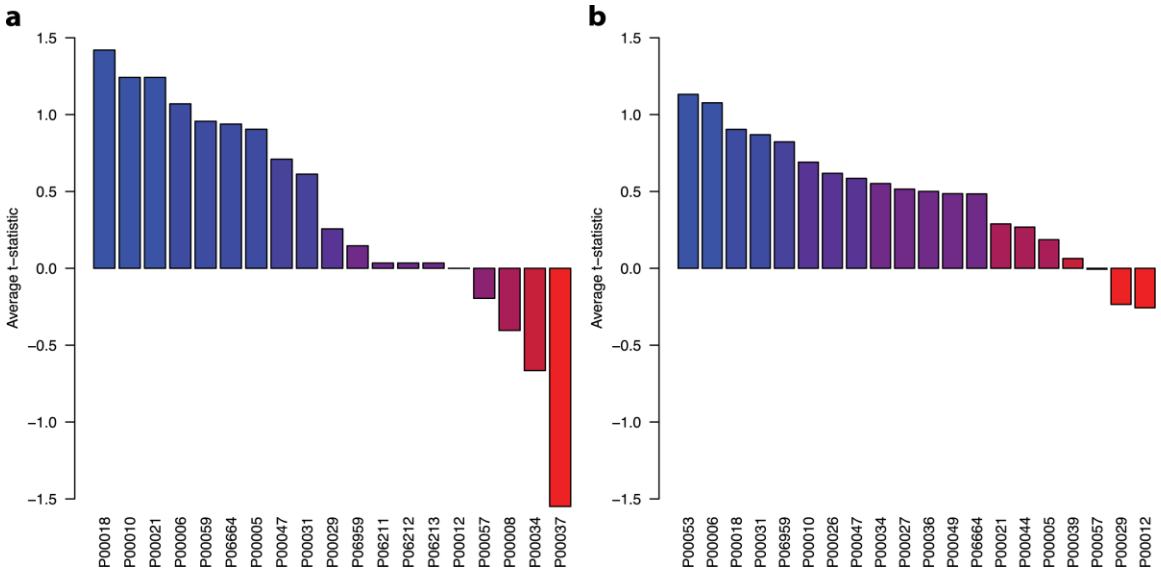

Figure 20 Average t-statistics for PANTHER pathways enriched in final models.

Plots show the average t-statistic for most prevalent pathways. Mean gene expression values for each gene in a given pathway were compared for non-responders vs responders. (A) 309 identified genes from 5-FU pan-cancer model across pathways with highest gene percentage. (B) 1158 identified genes from GCB pan-cancer model across pathways with highest gene percentage.

5.4 Discussion

In our study we have shown that primary tumor gene expression can be a good predictor of cancer drug response. By utilizing different clustering and classification methods we predicted cancer drug response with model validation accuracy of up to 86%.

Our model validation results show stronger performance for the GCB model than the 5-FU model. We attribute the GCB model's high prediction accuracy to multiple facets. First, the GCB model had a more substantial sample size of 92 patients. This was undoubtedly beneficial, as the algorithm was able to take advantage of the increased diversity in the training data to build a more robust model. Secondly, when clustering the

gene expression levels using a similarity threshold, the GCB data was grouped into fewer clusters; an indication the dataset was more homogenous and adaptable to dimension reduction. Further, random forest cross-validation accuracy was optimized at a higher number of clusters than 5-FU (73% more genes), leading us to believe the larger sample size helped the model to better differentiate informative features from noise. Another important variable between models is the specific cancer type on which model validation was performed. Our 5-FU training model was only able to predict with 76% accuracy on STAD patients (see

Table 45) compared to the 84.1% overall accuracy. This provides evidence that the pan-cancer model was better at predicting drug response for some cancer types than others.

We infer from our results that some drugs target mechanisms that are shared across a majority of cancers, while others may target mechanisms specific to certain cancer families. Our GCB pan-cancer model predicts all cancer types at comparable levels to that of the overall accuracy. On the other hand, a much higher variation in accuracy is seen from the 5-FU model (

Table 45). In the cases where the targeted mechanisms of a drug are different across cancer families, we would expect to see a reduction in the prediction accuracies of cancers with dissimilar mechanisms. When more data becomes available, future work could test the performance of models built on molecularly similar cancers of the same histology or anatomy, as suggested by a recent study [255].

Our gene set enrichment analysis reveals that many biological pathways relevant to drug metabolism and cancer are present in our most predictive gene modules. Both 5-FU

and GCB pan-cancer models' predictors were found to have a high percent of genes from the *WNT* signaling pathway, 11.8% and 5.7%, respectively. This pathway's contribution to tumorigenesis via cell fate determination and cell migration have already been proven [256]. Moreover, upregulation of the *WNT* pathway is involved in more than 30% of gastric cancer cases [257]. This is supported by our data since stomach adenocarcinoma is the most populous cancer type in the 5-FU pan-cancer model (Table 7). Both pan-cancer models were also enriched for genes from the "inflammation mediated by chemokine and cytokine signaling" pathway. Chemokines direct trafficking and migration of immune cells and inhibition of these proteins has been proven effective in preventing the accumulation of leukocytes near sites of inflammation [254]. Lastly, integrin signaling was found within the top seven pathways from

Table 8. Integrins are adhesion receptors that allow cells to respond to signals from the surrounding microenvironment by interacting with the extracellular matrix [258]. They have been implicated in cell adhesion-mediated drug resistance, a pro-survival and anti-apoptotic function [259]. We believe the presence of these pathways in our models provides insight into their biological relevance and tactic for predicting cancer drug response.

5.5 Conclusions

The results of our final approach for prediction cancer drug responses were conclusively accurate and, more importantly, interpretable. Our classifier selected genes that are integral parts of drug metabolism and cancer biology. This combination of

accuracy and interpretability has been difficult to achieve in predictive models attempted in the past. We attribute our success to the utilization of in-vivo gene expression data, which eliminates the need to extrapolate human drug responses from cell-line or other in-vitro features. Furthermore, our implementation of optCluster and random forest provided us with a method to perform dimension reduction in a biologically informative manner. Feature ranking, as we have shown, selects biologically relevant genes, that may yield new therapeutic targets. While recent discussion has suggested machine learning can appear as "alchemy" [260], we encourage the continued effort in the field of personalized cancer medicine as it bears great potential for benefiting patients. To conclude, predicting cancer drug response from patient RNA-seq data will be an important tool for personalized oncology. We anticipate that predictive models, such as the ones we present, will continue to grow more powerful and will provide clinicians and patients with additional information to aid in selecting first or second-line therapies.

5.6 Methods

5.6.1 Clustering & Variable Selection

Based on results from a previous study [25], which showed pan-cancer models out-performed single cancer models, along with our own empirical analysis (see Supplementary Methods), our study focused on pan-cancer models. We relied on clustering methods for dimension reduction. We implemented six clustering algorithms (clara, hierarchical, k-means, model, pam, and sota) [261-265] to determine which would provide

the highest percentage of correct predictions (prediction accuracy). We selected the top 5,000 genes for clustering based on gene expression variability (Figure 43).

To determine accuracy, we used the mean value for each gene module (cluster assignments) as a variable for prediction. The number of observations were lower than the chosen number of clusters, so variable selection was performed. We used random forest to determine variable importance. Random forest utilizes a forest of binary trees to split the data into multiple subsets based on predictive power. To stabilize the ranked list of variable importance, we took the mean Gini value for each gene module after 200 random forest runs. Mean decrease Gini was selected as it helped control for overfitting, an issue we were conscious of given the small sample size. We determined the optimal number of variables for prediction by ranging the number of variables from one to the number of clusters and using random forest a second time to classify patients as responder or non-responder.

The accuracy obtained from the six clustering algorithms were computed and compared. We also attempted clustering with +/- 50% of the starting number of genes, but these yielded worse performances. For the analysis, we used the R package, optCluster, and relied on internal validation to determine the stability of the clusters [252]. For each clustering method, we determined the best model by exploring the impact of the number of clusters on prediction accuracy.

To find the best clustering algorithm for our study, we performed clustering and classification in a five-step process (Figure 44).

1. The algorithm was run with *N* number of clusters (where N is in the range of 180-

220; accuracies were observed to drop outside this cluster number range).

2. Random forest was used to determine variable importance and perform variable selection.
3. For each top *n* most important variables (where *n* is between 1 to *N*), random forest was used to classify each patient.
4. The highest classification method was logged for the best value of *N* and *n* for each clustering method and for each drug model.
5. The best model from the previous step was selected and the parameters were tuned to capture any additional accuracy.

To further validate the results, we assessed the accuracy rates of logistic regression and support vector machines when applied on our optimal random forest features. We also tested the impact of demographic data of the patients (gender, age, cancer type and cancer stage) on the accuracy of the model.

5.6.2 Model Validation

The final 5-FU and GCB models were trained on their respective pan-cancer data in which half of the patients from the most populous cancer, STAD or PAAD respectively, were held out as the validation sets. To reduce the impact of small sample size in the pan-cancer models, we used bootstrap as an up-sampling method to increase the size of the training set by 50%. The two validation sets consisted of 17 randomly selected STAD patients and 28 randomly selected PAAD patients for the 5-FU and GCB model respectively.

5.6.3 Statistical Methods: Gene Enrichment Analysis

Our hypothesis was that the selected genes in our final model would be enriched for pathways involved in drug metabolism or cell signaling. To test this hypothesis, we performed two statistical tests. First, we performed a PANTHER overrepresentation test. Using the results of the PANTHER analysis, we performed a t-test comparing the mean gene expression of responders and non-responders for the top twenty pathways based on the percent of genes hit against the total number of genes in the pathway.

APPENDIX A.

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

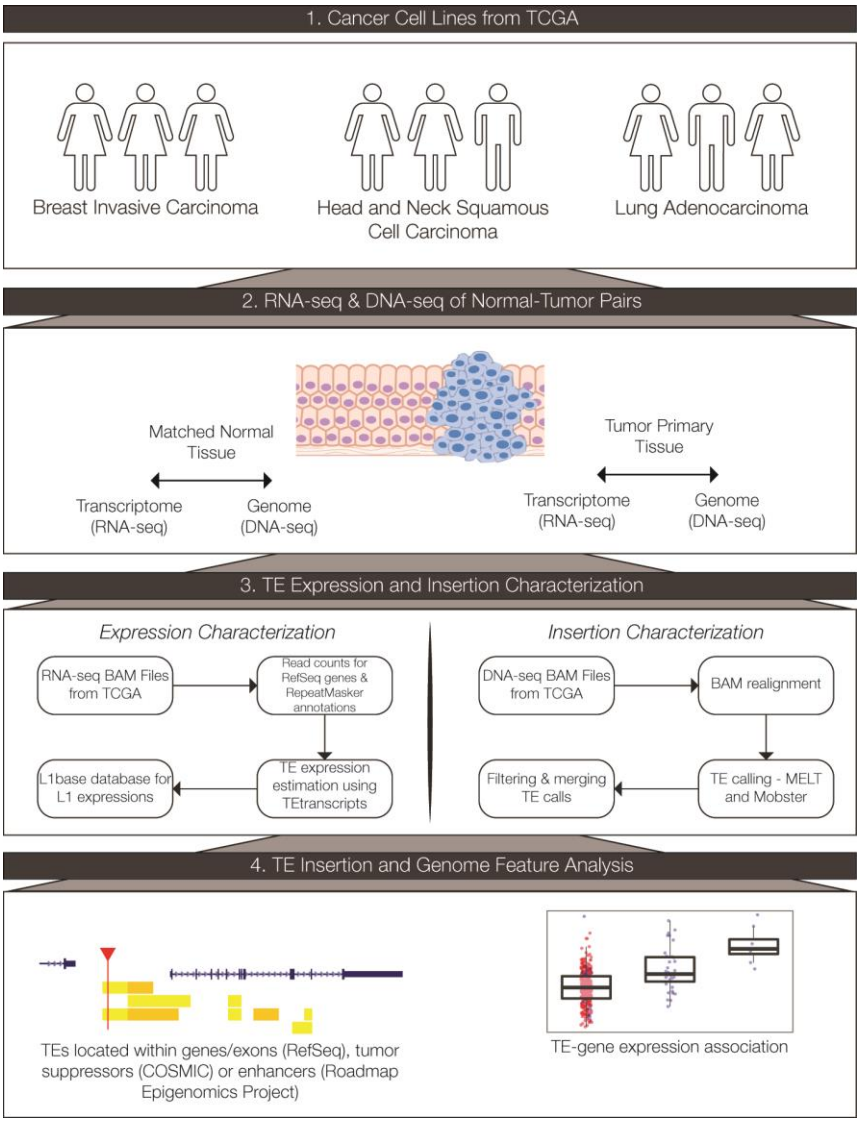

Figure 21 Scheme of the analysis pipeline used for this study.

1) Matched normal and primary tumor sample data for three patients each from three cancer types were obtained from TCGA. 2) Transcriptome (RNA-seq) and whole genome sequence (DNA-seq) data were compared for normal versus tumor tissue samples. 3) RNA-seq and DNA-seq data were analyzed to characterize TE expression levels and TE insertional activity for normal versus tumor tissue samples as shown. 4) Genomic features associated with tumor-specific TE insertions were evaluated to look for putative TE cancer causing mutations.

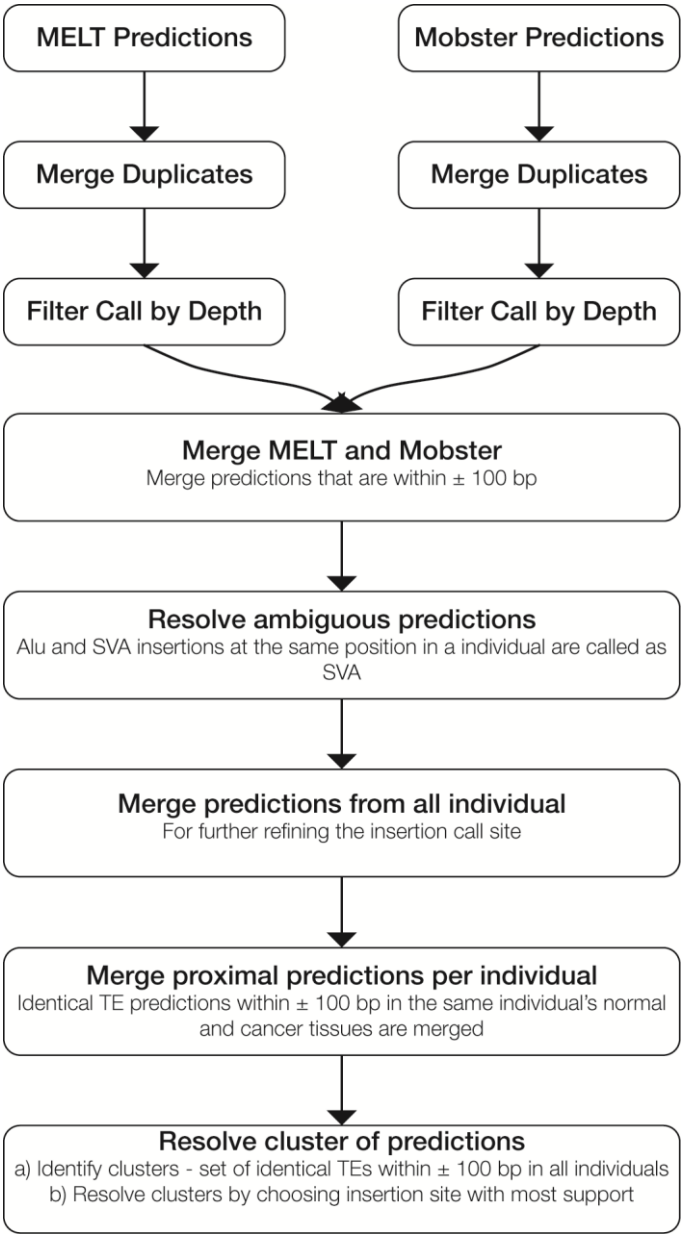

Figure 22 Scheme of the TE insertion detection analysis pipeline used for this study.

Steps used to merge predictions from the MELT and Mobster programs are shown along with the post-processing steps used to ensure that accurate TE insertion predictions were chosen for subsequent analysis.

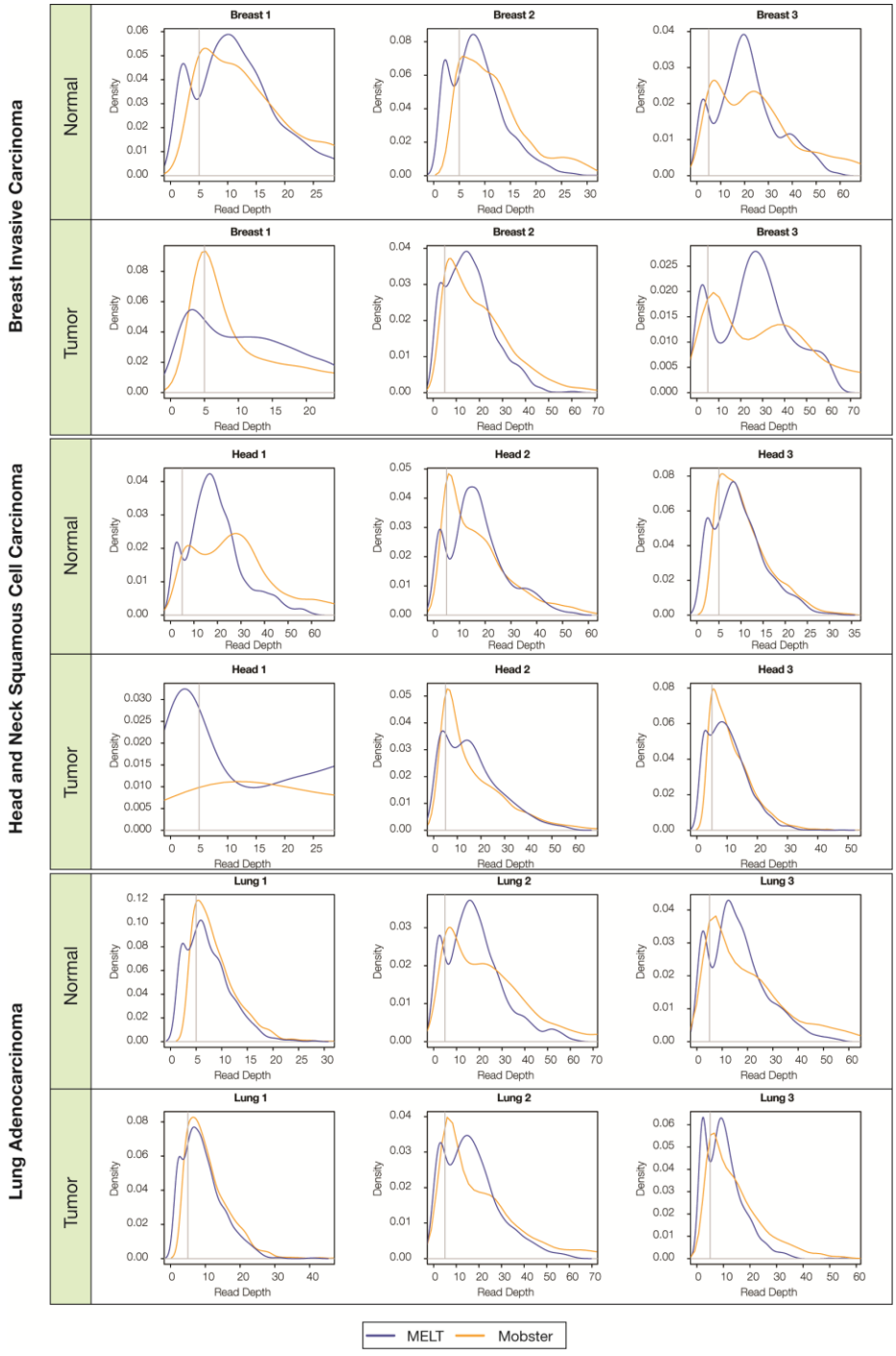

Figure 23 Density distributions for the numbers of mapped reads supporting TE insertion calls.

Read depth distributions are shown for TE insertion calls made with the MELT (blue) and Mobster (orange) programs for all 18 of the matched normal and primary tumor tissue samples analyzed here. The locations of the lower read depth threshold of 5 reads are indicated for each distribution with a gray line, and the distributions are all bounded by the upper read depth threshold corresponding to 4X the average sequencing depth of the sample.

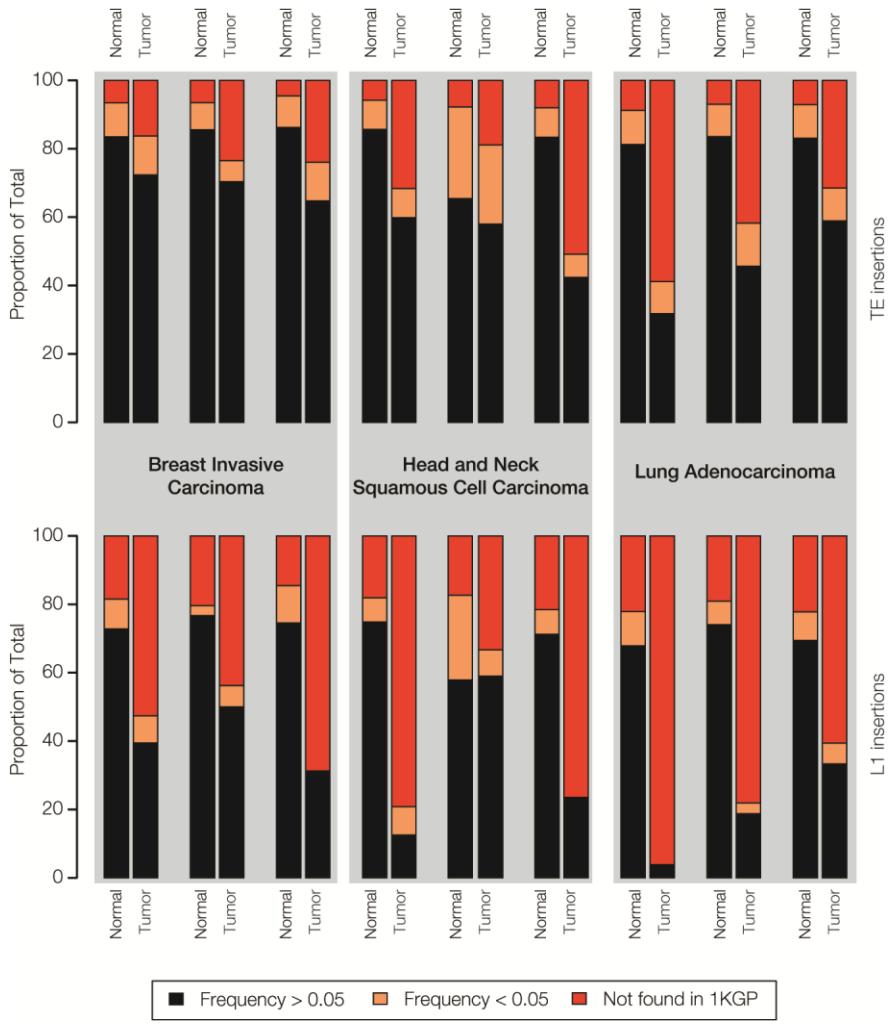

Figure 24 Population frequencies of observed TE insertions in matched normal versus tumor tissue pairs are shown for all of the TEs analyzed here and for L1s alone.

Frequencies are represented as in Figure 4, but data are shown for each individual sample across the three cancer types analyzed here.

APPENDIX B.

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

Table 10 Data sources, programs, and statistical methods used in this study.

Name	Description	Reference
Data Sources		
TCGA	RNA-seq data from matched normal-tumor patient samples from 13 cancer types	[166]
COSMIC	Cancer Gene Census (CGC) annotations for 723 cancer-associated genes	[266]
RepeatMasker	Genomic coordinates and annotations for human TEs	[129]
NCBI RefSeq	Genomic coordinates (exon/intron boundaries) for human genes	[127]
GENCODE	Genomic coordinates (exon/intron boundaries) for human genes	[267]
Genomic Data Commons (GDC)	Coordinates and event counts of alternative splicing events in TCGA samples	[122]
Programs		
SplAdder	Detection and quantification of alternative splicing events	[268]
BEDTools	Identification of alternative splicing events colocated near TE sequences	[71]
DESeq2	Normalization of alternative splice isoform expression across patients within a cancer type	[131]
UCSC Genome Browser	Visualization of putative TE-derived isoforms in cancer associated genes	[269]
Statistical Methods		
Variance Stabilizing Transformation (VST)	Blind transformation used to remove the experiment-wide trend of variance over mean, normalize alternative splice isoform expression values and produce log2 scale data	[131]
Single linkage clustering	Algorithm to merge overlapping alternative splice isoforms based on $\geq$75% overlap of isoform genomic coordinates in an agglomerative fashion	
Relative expression change (REC)	Quantification of the normalized change in expression levels of TE-derived alternative splice isoforms in tumor versus normal tissue	
G-test	Maximum likelihood statistical significance test for 2x2 isoform expression contingency matrix	

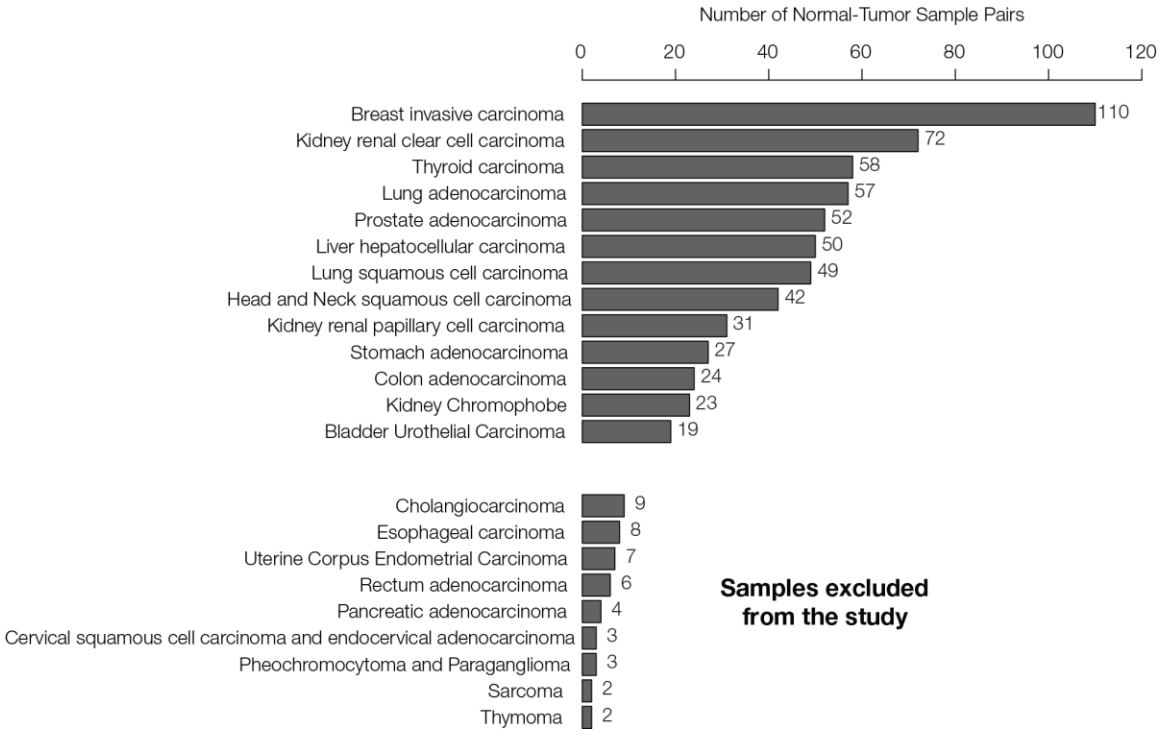

Figure 25 Number of patient samples per cancer type analyzed here.

RNA-seq data for matched normal-tumor sample pairs were taken from The Cancer Genome Atlas (TCGA). Cancers with less than 10 sample pairs were excluded from further analysis.

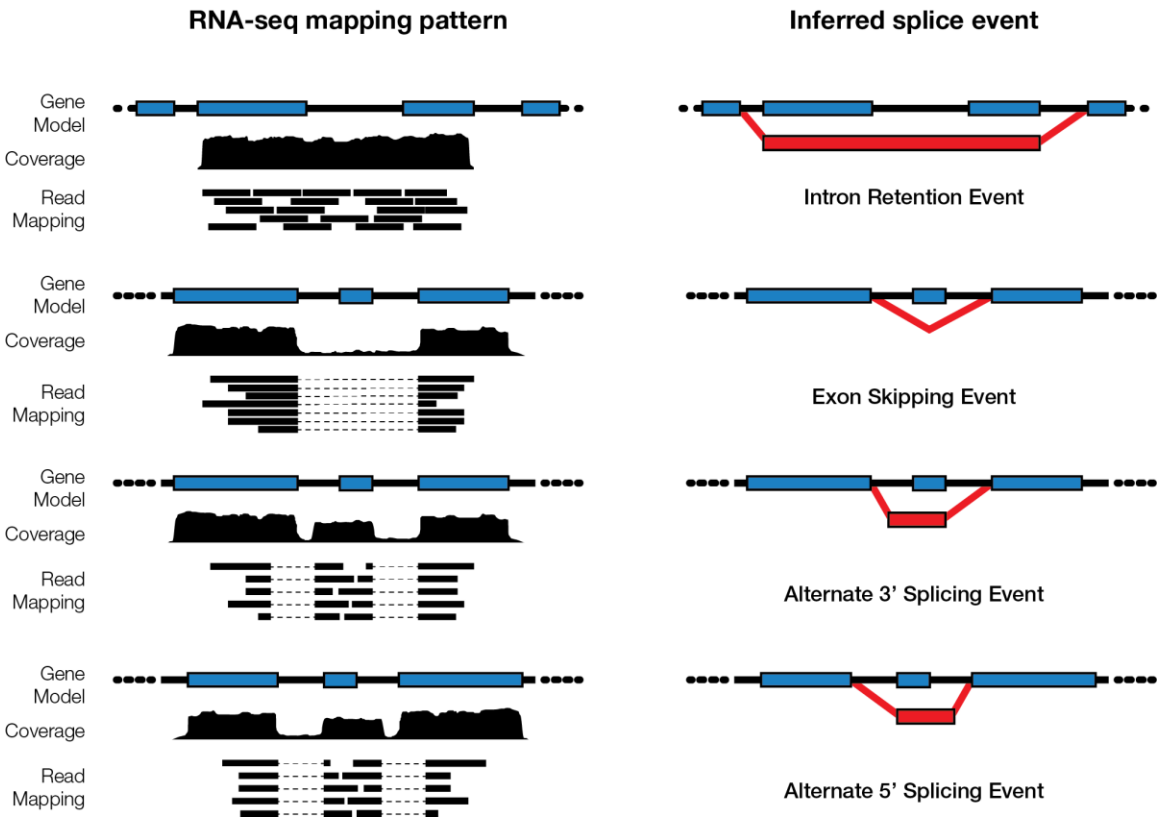

Figure 26 Alternative splicing event types analyzed here.

Four kinds of alternative splice events were analyzed for this study: intron retention, exon skipping, alternate 3' splicing, and alternate 5' splicing. Splicing events were identified and characterized based on the mapping of RNA-seq reads to gene models, using the program SplAdder as previously described [268]. *For each type of splicing event, its corresponding RNA-seq read mapping pattern is shown adjacent to a schematic of the inferred splicing event type.*

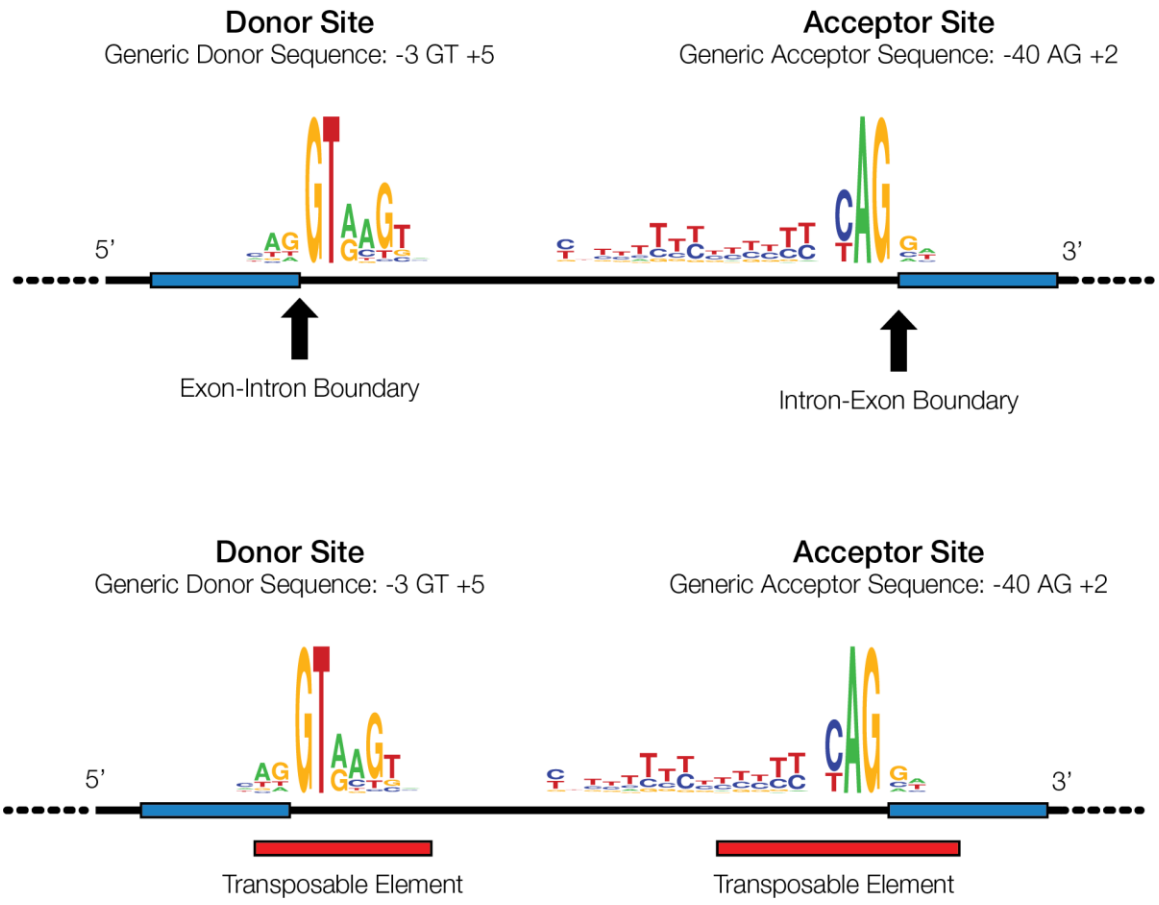

Figure 27 Scheme for the identification TE-derived splice sites.

The top panel shows 3' and 5' exon boundaries along with their canonical splice donor and acceptor site sequence motifs [242]. *Potential TE-derived splice donor and acceptor sites were identified where TE sequences were found to overlap the canonical splice site motifs as shown in the bottom panel.*

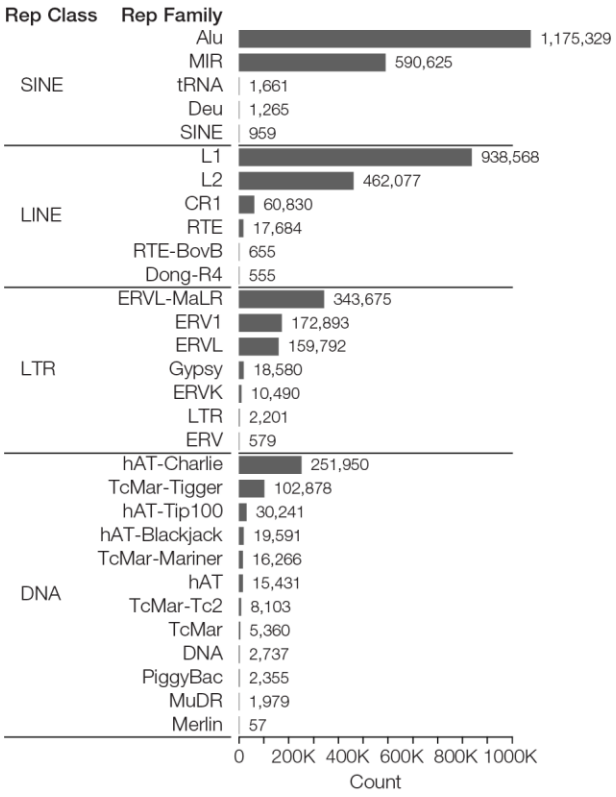

Figure 28 Counts of human transposable element (TE) sequences in the human genome.

TE names and counts are taken from RepeatMakser annotations. TEs are grouped into four major classes, and TE family names are shown for each class. The four major classes are: SINE – short interspersed nuclear element, LINE – long interspersed nuclear element, LTR – long terminal repeat containing element, and DNA – DNA-type element. SINEs, LINEs, and LTRs are retrotransposons that transpose via a copy and paste mechanism catalyzed by reverse transcriptase; DNA-type elements transpose via a cut and paste mechanism catalyzed the transposase enzyme.

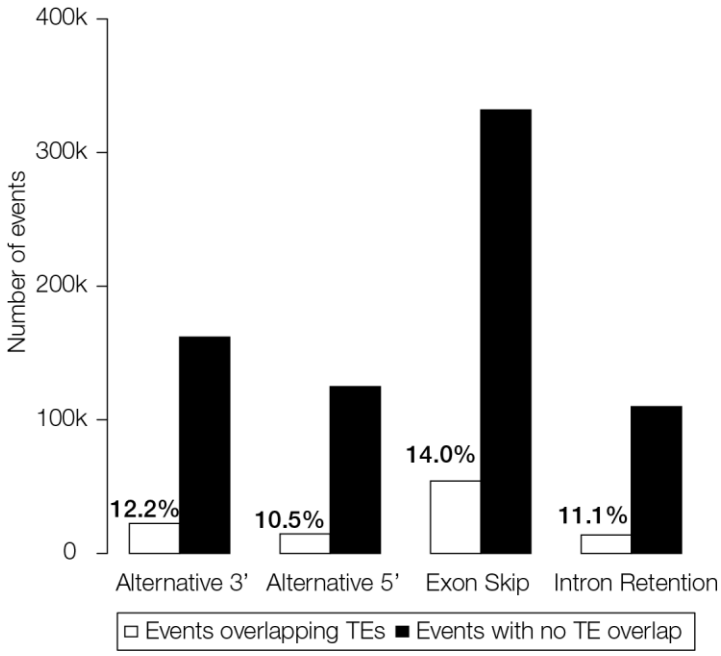

Figure 29 Number of alternative splice events seen for human genes.

Counts for the four different alternative splice event types are shown for TE-derived (white) versus non TE-derived isoforms (black). The percentages of TE-derived events are shown.

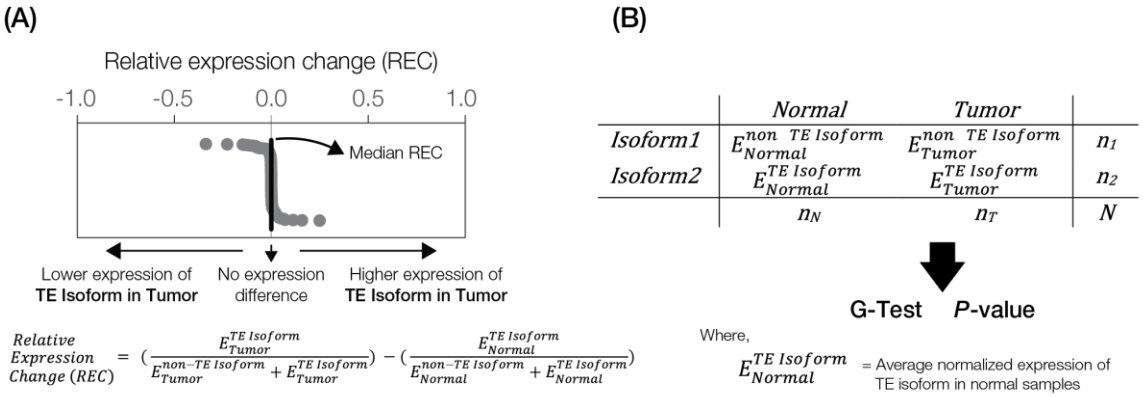

Figure 30 Quantification and statistical testing for differential expression of TE-derived alternative splice events.

(A) The relative Expression Change (REC) metric quantifies the normalized change in expression levels of TE-derived alternative splice isoforms in tumor versus normal tissue. This metric accounts for the expression of TE and non-TE isoform in both normal and tumor issues. Higher REC values indicate relatively higher expression of TE isoform in tumor tissue and vice versa. Details on the expression counts and formulas can be found in the Methods section. (B) Formulation of the 2x2 contingency matrix used for the G-test of the significance of expression difference.

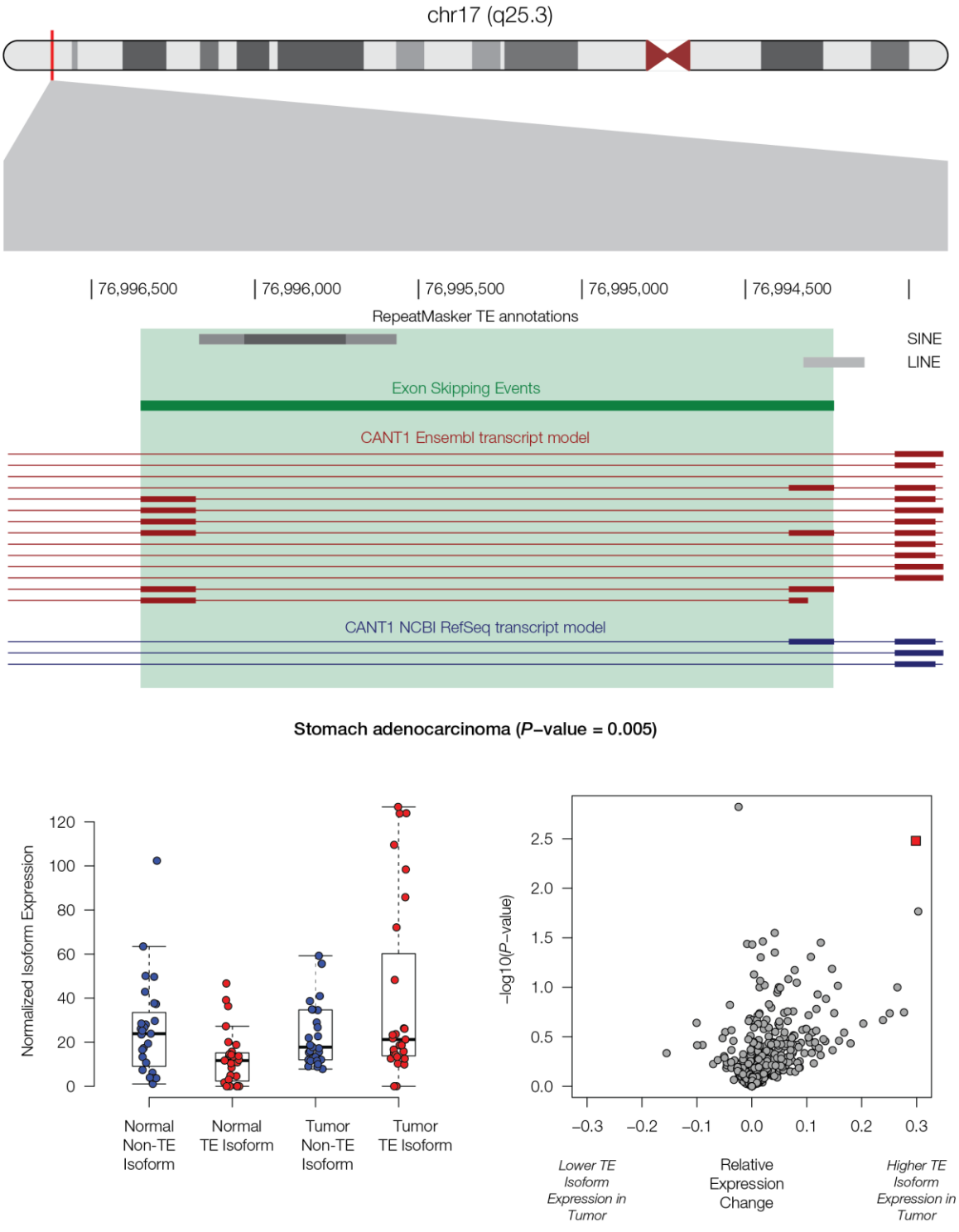

Figure 31 TE-derived alternative splicing in the CANT1 gene.

(A) The location of CANT1 on the long arm of chromosome 17 is shown along with the specific location of its TE-derived alternative splicing event. The presence of LINE and SINE sequences result in an exon skipping event. (B) Distributions of the non-TE (blue)

and TE-derived (red) isoforms are shown for matched normal (left) and stomach adenocarcinoma samples (right). (C) Relative expression change (REC) values are plotted against the corresponding G-test P-values (see Methods and Figure 30) for the matched normal and stomach adenocarcinoma samples. The CANT1 TE-derived isoform values are shown as a red square.

APPENDIX C.

SUPPLEMENTARY INFORMATION FOR CHAPTER 4

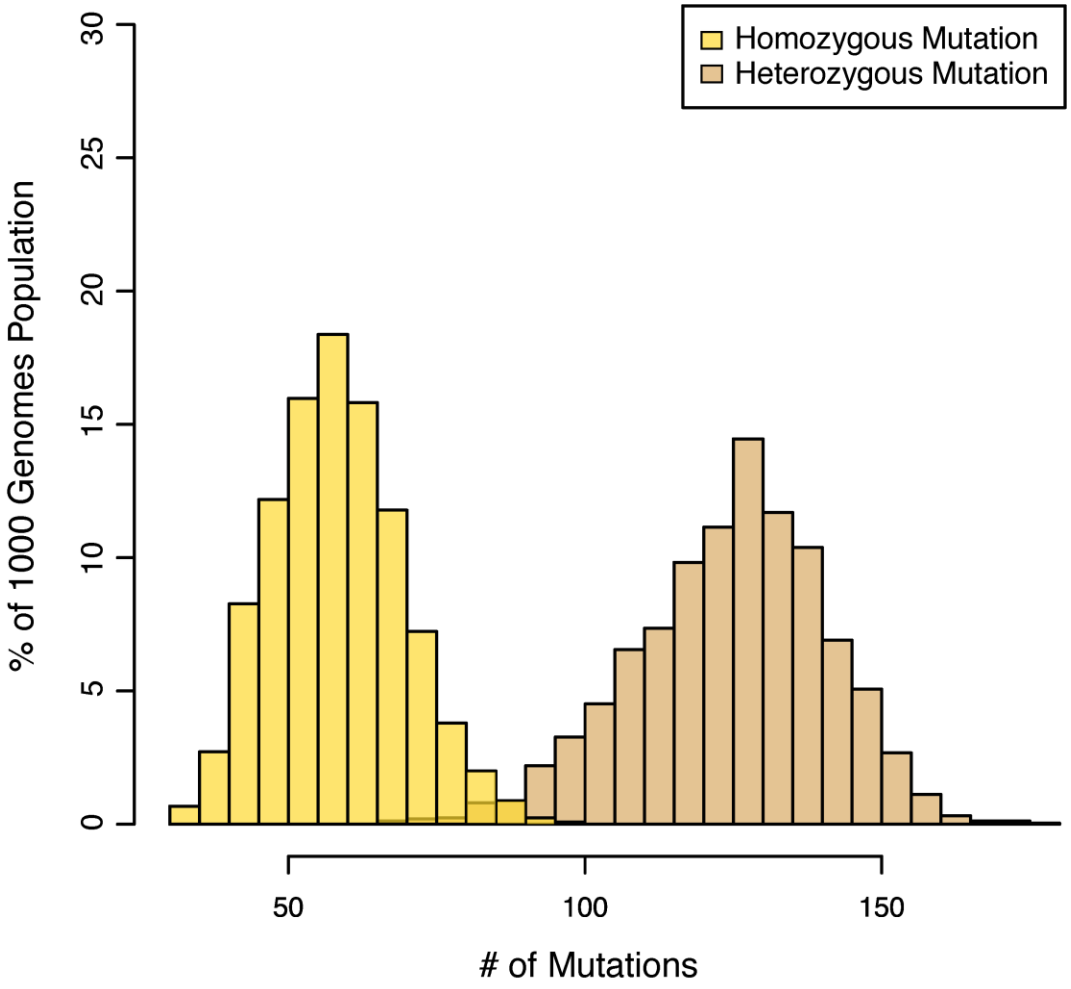

Figure 32 Distribution of COSMIC census mutations in the 1KGP.

The distribution of cancer associated mutations (all coding mutations in COSMIC Census Genes) within the 1KGP, color-coded by homozygous and heterozygous mutation as shown in the key.

Table 11 448 LoF COSMIC census mutations in TSGs of 1KGP.

cosmicIds	GeneName	Description	RoleInCancer	Hets	Homs	1KG_AF
COSM44637	*TP53*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM6458552	*APOBEC3B*	Deletion - Frameshift	oncogene, TSG	87	4	0.0189696
COSM4766174	*TET2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4040370	*POLE*	Substitution - Missense	TSG	2	0	3.99E-04
COSM6357161	*ARNT*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM912247	*CAMTA1*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM302197	*FANCA*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3692861	*CBLC*	Substitution - Missense	oncogene, TSG	62	1	0.0127796
COSM96955	*SUFU*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3528373	*TMED1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3152880	*SGSM3*	Substitution - Missense	oncogene, TSG, fusion	3	0	5.99E-04
COSM2152381	*PPARG*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM1738822	*TP63*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM1740190	*CASC5*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM898756	*SPEN*	Substitution - Missense	TSG	1	0	2.00E-04
COSM97174	*PRDM1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1709781	*PER1*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM1222525	*PTPN13*	Substitution - Missense	TSG	1	0	2.00E-04
COSM13713	*RP11-145E5.5*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3279826	*PALB2*	Substitution - Missense	TSG	4	0	7.99E-04
COSM4707624	*NOTCH2*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM1471274	*ERBB4*	Substitution - Nonsense	oncogene, TSG	1	0	2.00E-04
COSM4525688	*FAT1*	Substitution - Missense	TSG	3	0	5.99E-04
COSM4852474	*KAT6B*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM1375475	*FES*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM6469471	*FAT1*	Substitution - Missense	TSG	17	0	0.00339457
COSM5945710	*ASXL1*	Substitution - Missense	TSG	4	0	7.99E-04
COSM3117280	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1051055	*FAT4*	Substitution - Missense	TSG	2	0	3.99E-04
COSM1321925	*ATM*	Substitution - Missense	TSG	5	0	9.98E-04
COSM861	*RB1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6002376	*ZFHX3*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1162368	*KMT2C*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1581248	*KMT2C*	Substitution - Missense	TSG	4	0	7.99E-04
COSM3665212	*AXIN2*	Substitution - Missense	TSG	2	0	3.99E-04
COSM3291839	*PTPRT*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2911748	*DNMT3A*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5979811	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	2	0	3.99E-04
COSM5981687	*ASXL1*	Substitution - Missense	TSG	2	0	3.99E-04
COSM403987	*BCOR*	Substitution - Missense	TSG, fusion	1	0	2.00E-04

Table 12 (continued).

COSM1035959	*CBLB*	Substitution - Missense	TSG	1	0	2.00E-04
COSM205536	*ATR*	Substitution - Missense	TSG	1	0	2.00E-04
COSM87161	*TET2*	Substitution - Missense	TSG	2	0	3.99E-04
COSM50512	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	3	0	5.99E-04
COSM1058511	*PTPN13*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5989825	*RAD21*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM87153	*TET2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1702652	*KAT6B*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM5792284	*BIRC3*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM925	*RB1*	Substitution - Missense	TSG	4	0	7.99E-04
COSM4555574	*TNFAIP3*	Substitution - Missense	TSG	1	0	2.00E-04
COSM13463	*RP11-145E5.5*	Substitution - Missense	TSG	11	0	0.00219649
COSM1172017	*NOTCH2*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM1015967	*ERBB4*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM29723	*MLH1*	Unknown	TSG	1	0	2.00E-04
COSM485085	*KMT2C*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM21828	*ATM*	Substitution - Missense	TSG	3	0	5.99E-04
COSM5602899	*BCL9L*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM6203207	*KMT2D*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM2135245	*RAD51B*	Substitution - Missense	TSG, fusion	9	0	0.00179712
COSM975877	*NCOR1*	Substitution - Missense	TSG	5	0	9.98E-04
COSM4990375	*N4BP2L1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6023789	*MLH1*	Substitution - Missense	TSG	6	0	0.00119808
COSM4171872	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM332031	*FAT4*	Substitution - Missense	TSG	9	0	0.00179712
COSM1197389	*FAT1*	Substitution - Missense	TSG	7	0	0.00139776
COSM3899739	*NRG1*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM5981818	*LZTR1*	Substitution - Missense	TSG	3	0	5.99E-04
COSM1241437	*KMT2C*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4015501	*PRF1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5940810	*SMAD2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5762836	*ASXL1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM289289	*MED12*	Substitution - Missense	TSG	1	0	2.00E-04
COSM241285	*PTCH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM368041	*NCOR2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM41644	*TET2*	Substitution - Nonsense	TSG	2	0	3.99E-04
COSM4682092	*EXT2*	Substitution - Missense	TSG	2	0	3.99E-04
COSM990606	*KEAP1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5945587	*PHF6*	Substitution - Missense	TSG	0	0	0
COSM3733340	*FAT4*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4698764	*LRIG3*	Substitution - Missense	TSG, fusion	6	0	0.00119808

Table 13 (continued).

COSM3084735	*PMS2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1709385	*ZFHX3*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM3284349	*TCF3*	Substitution - Missense	oncogene, TSG, fusion	3	0	5.99E-04
COSM3732749	*LRP1B*	Substitution - Missense	TSG	2	0	3.99E-04
COSM3967980	*SH2B3*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5376647	*BRCA2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4383892	*TET2*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM36249	*TNFAIP3*	Substitution - Missense	TSG	63	1	0.0129792
COSM6025620	*PML*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM1033838	*MYH9*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM4830573	*SLC34A2*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM186809	*MLH1*	Substitution - Missense	TSG	3	0	5.99E-04
COSM5020249	*TET2*	Substitution - Missense	TSG	383	40	0.0924521
COSM1054236	*FAT1*	Substitution - Missense	TSG	4	0	7.99E-04
COSM6469473	*FAT1*	Substitution - Missense	TSG	3	0	5.99E-04
COSM3417538	*N4BP2L1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3740582	*NOTCH2*	Substitution - Missense	oncogene, TSG	2	0	3.99E-04
COSM327337	*TET2*	Substitution - Missense	TSG	12	0	0.00239617
COSM99653	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	2	0	3.99E-04
COSM279944	*CLTCL1*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM404245	*CD274*	Substitution - Missense	TSG, fusion	3	0	5.99E-04
COSM224982	*EXT2*	Substitution - Missense	TSG	2	0	3.99E-04
COSM1006985	*LRP1B*	Substitution - Missense	TSG	9	1	0.00219649
COSM1637799	*FHIT*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM2796667	*AXIN2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1035875	*VHL*	Substitution - Missense	TSG	1	0	2.00E-04
COSM42084	*TET2*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM5357611	*ARID1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1351719	*BCL9L*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM142703	*PTPRT*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1714395	*APOBEC3B*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM4039331	*ETV6*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM6329722	*POLE*	Substitution - Missense	TSG	1	0	2.00E-04
COSM947808	*RB1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4139978	*PER1*	Substitution - Missense	TSG, fusion	16	0	0.00319489
COSM4989668	*CBLC*	Substitution - Missense	oncogene, TSG	3	0	5.99E-04
COSM5020556	*PTPN13*	Substitution - Missense	TSG	1	0	2.00E-04
COSM303861	*RP11-145E5.5*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6413439	*CBFA2T3*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM278449	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4596715	*FAT4*	Substitution - Missense	TSG	11	0	0.00219649

Table 14 (continued).

COSM5712841	*TCF3*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM6229984	*NTRK1*	Substitution - Missense	oncogene, TSG, fusion	12	0	0.00239617
COSM1050869	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1581234	*KMT2C*	Substitution - Missense	TSG	3	0	5.99E-04
COSM3639951	*PMS2*	Substitution - Missense	TSG	5	0	9.98E-04
COSM4733147	*TET2*	Substitution - Missense	TSG	2	0	3.99E-04
COSM5020081	*TYW1*	Unknown	TSG	8	0	0.00159744
COSM4126516	*PTPN13*	Substitution - Missense	TSG	2	0	3.99E-04
COSM5979034	*CTC-554D6.1*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4040376	*POLE*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3505442	*FES*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM3291831	*PTPRT*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3687129	*ARHGEF12*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM1195213	*ETNK1*	Substitution - Missense	TSG	2	0	3.99E-04
COSM6289411	*CHEK2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3793095	*ERCC5*	Substitution - Missense	TSG	21	1	0.00459265
COSM6023913	*CDH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5576878	*NOTCH2*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM5748604	*NTRK1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM4928397	*CLTCL1*	Substitution - Missense	TSG, fusion	3	0	5.99E-04
COSM4799369	*DDX10*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM4415590	*FANCA*	Substitution - Missense	TSG	279	37	0.0704872
COSM1087645	*KMT2C*	Substitution - Missense	TSG	1	0	2.00E-04
COSM245482	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM21637	*ATM*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5462513	*PKD1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5507365	*TERT*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM3042197	*ERCC3*	Substitution - Missense	TSG	13	0	0.00259585
COSM1408311	*MSH6*	Substitution - Missense	TSG	2	0	3.99E-04
COSM6035524	*ARHGEF12*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM1221785	*PPP2R1A*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4899601	*PALB2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6466277	*LRP1B*	Substitution - Missense	TSG	24	0	0.00479233
COSM14367	*VHL*	Substitution - Missense	TSG	2	0	3.99E-04
COSM1496821	*WRN*	Substitution - Missense	TSG	2	0	3.99E-04
COSM5020362	*SUFU*	Substitution - Missense	TSG	3	0	5.99E-04
COSM3672430	*CDK12*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4538895	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3968227	*ARID2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4748966	*CREBBP*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM4665561	*BCORL1*	Substitution - Nonsense	oncogene, TSG	1	0	2.00E-04

Table 15 (continued).

COSM1256743	*LRP1B*	Substitution - Missense	TSG	10	0	0.00199681
COSM1039402	*ATR*	Substitution - Missense	TSG	1	0	2.00E-04
COSM248836	*APC*	Substitution - Missense	TSG	0	0	0
COSM262336	*NCOR1*	Substitution - Missense	TSG	12	0	0.00239617
COSM5990008	*CDH11*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM4672963	*CLTCL1*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM3591209	*TP63*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM3009860	*FAT4*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4720948	*RECQL4*	Substitution - Missense	oncogene, TSG	3	0	5.99E-04
COSM3968387	*WIF1*	Substitution - Nonsense	TSG, fusion	1	0	2.00E-04
COSM93978	*BRCA1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4749447	*TP53*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM4606547	*SETD2*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM5879023	*TET2*	Unknown	TSG	1	0	2.00E-04
COSM3740792	*SPEN*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5021138	*CLTCL1*	Substitution - Missense	TSG, fusion	5	0	9.98E-04
COSM5899948	*PBRM1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1716814	*PTPN13*	Substitution - Missense	TSG	40	1	0.00838658
COSM1058526	*PTPN13*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1216588	*NAB2*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM3364915	*VHL*	Substitution - Missense	TSG	5	0	9.98E-04
COSM3392822	*FAT1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2871917	*RECQL4*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM22487	*ATM*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4589981	*LZTR1*	Unknown	TSG	893	2	0.179113
COSM5989823	*FAT4*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4683689	*FAT4*	Substitution - Missense	TSG	4	0	7.99E-04
COSM133736	*DNMT3A*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM3724370	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	2	0	3.99E-04
COSM1471496	*EP300*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM87012	*DNMT3A*	Substitution - Missense	TSG	0	0	0
COSM473150	*BRIP1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2871931	*RECQL4*	Substitution - Missense	oncogene, TSG	2	0	3.99E-04
COSM6338144	*BCL9L*	Substitution - Missense	oncogene, TSG	2	0	3.99E-04
COSM1297701	*BCL9L*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM231547	*DNMT3A*	Substitution - Missense	TSG	1	0	2.00E-04
COSM42076	*TET2*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM3131780	*FAT1*	Substitution - Missense	TSG	2	0	3.99E-04
COSM5494125	*PTK6*	Substitution - Missense	oncogene, TSG	4	0	7.99E-04
COSM4893689	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1204277	*DNM2*	Substitution - Missense	TSG	1	0	2.00E-04

Table 16 (continued).

COSM5985274	*KMT2C*	Substitution - Missense	TSG	14	0	0.00279553
COSM1679262	*ZFHX3*	Substitution - Missense	TSG	2	0	3.99E-04
COSM2739084	*NCOR1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1194722	*FAT1*	Substitution - Missense	TSG	5	0	9.98E-04
COSM21938	*ATM*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4383801	*TET2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1050971	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3699056	*NRG1*	Substitution - Missense	TSG, fusion	167	15	0.0393371
COSM922707	*ATM*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4382532	*BMPR1A*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM6461750	*CARS*	Substitution - Missense	TSG, fusion	10	0	0.00199681
COSM6324495	*FAT1*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM3443164	*ATM*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3100110	*ELL*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM986144	*PER1*	Substitution - Missense	TSG, fusion	3	0	5.99E-04
COSM20674	*NTRK1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM5428669	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5764815	*PBRM1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5980437	*POLE*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3600130	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1058546	*PTPN13*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4384332	*KMT2C*	Substitution - Missense	TSG	1	0	2.00E-04
COSM947376	*FOXO1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM4129165	*CDH11*	Substitution - Missense	TSG, fusion	673	118	0.18151
COSM3520127	*RNF43*	Substitution - Missense	TSG	10	0	0.00199681
COSM4103837	*MYH9*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM3559467	*ATP2B3*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1243234	*EXT2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3279882	*DCTN5*	Substitution - Missense	TSG	3	0	5.99E-04
COSM601786	*PMS2*	Substitution - Missense	TSG	16	0	0.00319489
COSM1184281	*AXIN2*	Substitution - Missense	TSG	4	0	7.99E-04
COSM6339935	*TGFBR2*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4018618	*ARHGEF12*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM4440896	*RUNX1T1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM43680	*TP53*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM5369878	*RP11-145E5.5*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3513236	*CBFA2T3*	Substitution - Missense	TSG, fusion	6	0	0.00119808
COSM3402556	*GRIN2A*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4419889	*DNM2*	Substitution - Missense	TSG	11	0	0.00219649
COSM4716121	*PMS2*	Substitution - Missense	TSG	22	0	0.00439297
COSM6229194	*KMT2C*	Substitution - Missense	TSG	1	0	2.00E-04

Table 17 (continued).

COSM5880093	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM4588060	*WRN*	Substitution - Missense	TSG	6	0	0.00119808
COSM5851889	*NCOR1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM11606	*TP53*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM4999323	*SLC25A1*	Deletion - Frameshift	TSG, fusion	20	1	0.00439297
COSM1032073	*CLTCL1*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM1642075	*PPARG*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM5870389	*ZBTB22*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM4584283	*MLH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1039312	*XPC*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5548996	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM1375971	*ERCC4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4980437	*ATP2B3*	Substitution - Missense	TSG	0	0	0
COSM3879563	*KMT2C*	Substitution - Missense	TSG	2	0	3.99E-04
COSM1199596	*CARS*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM2226965	*NCOR2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM144596	*KMT2D*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM4986919	*PALB2*	Substitution - Missense	TSG	43	0	0.00858626
COSM211728	*TET2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1718902	*FOXO3*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM5980256	*CDKN1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1394499;COSM1683551	*CBLC*	Insertion - Frameshift	oncogene, TSG	360	37	0.0866613
COSM4383607	*DNMT3A*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM4923010	*DICER1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5019704	*BRCA2*	Substitution - Missense	TSG	6	0	0.00119808
COSM4059146	*TSC2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4406524	*SPEN*	Substitution - Missense	TSG	1	0	2.00E-04
COSM719370	*ERBB4*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM4416035	*MSH6*	Substitution - Missense	TSG	14	0	0.00279553
COSM41741	*TET2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5008355	*FAT4*	Substitution - Missense	TSG	44	2	0.00958466
COSM5020989	*ATM*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4039895	*NCOR2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5940326	*PTPRB*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM3390954	*LRP1B*	Substitution - Missense	TSG	2	0	3.99E-04
COSM1691126	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5341968	*POLE*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3765996	*NCOR1*	Substitution - Missense	TSG	20	0	0.00399361
COSM3151672	*APOBEC3B*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM5765556	*LEF1*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM5749277	*PTPRB*	Substitution - Missense	TSG	1	0	2.00E-04

Table 18 (continued).

COSM5985218	*FAT1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4988714	*SDHB*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5515797	*MKL1*	Substitution - Missense	oncogene, TSG, fusion	2	0	3.99E-04
COSM5711936	*PTPN13*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5019598	*LRP1B*	Substitution - Missense	TSG	20	0	0.00399361
COSM4119165	*PBRM1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2813147	*DNM2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4698822	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM441184	*LRP1B*	Substitution - Missense	TSG	2	0	3.99E-04
COSM5019975	*NTRK1*	Substitution - Missense	oncogene, TSG, fusion	108	7	0.024361
COSM1212595	*KLF6*	Substitution - Missense	TSG	2	0	3.99E-04
COSM1232948	*XPC*	Substitution - Missense	TSG	6	1	0.00159744
COSM197777	*PTPN13*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5019884	*CLTCL1*	Substitution - Missense	TSG, fusion	40	1	0.00838658
COSM471318	*ERCC4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5042148	*MKL1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM4575243	*POLE*	Substitution - Missense	TSG	2	0	3.99E-04
COSM181752	*SPEN*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4992716	*ZBTB22*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM1222582	*PTPRB*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6412981	*ELL*	Substitution - Missense	TSG, fusion	3	0	5.99E-04
COSM36205	*ASXL1*	Substitution - Missense	TSG	26	0	0.00519169
COSM3152897	*MKL1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM6284223	*EPAS1*	Substitution - Missense	oncogene, TSG	2	0	3.99E-04
COSM1235467	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM5020963	*ATM*	Substitution - Missense	TSG	7	0	0.00139776
COSM3009627	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2871902	*RECQL4*	Substitution - Missense	oncogene, TSG	2	0	3.99E-04
COSM145584	*LRIG3*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM923664	*BCL9L*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM6475880	*TGFBR2*	Unknown	TSG	1	0	2.00E-04
COSM1321166	*BMPR1A*	Substitution - Missense	oncogene, TSG	5	0	9.98E-04
COSM13385	*MSH6*	Substitution - Missense	TSG	1	0	2.00E-04
COSM446120	*MAP3K13*	Substitution - Missense	oncogene, TSG	2	0	3.99E-04
COSM5920788	*BLM*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6352341	*CDH1*	Substitution - Missense	TSG	6	0	0.00119808
COSM5967258	*CHEK2*	Deletion - Frameshift	TSG	5	0	9.98E-04
COSM220521	*EP300*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM3186044	*MSH6*	Insertion - Frameshift	TSG	40	0	0.00798722
COSM5020928	*HPDL*	Substitution - Missense	TSG	12	0	0.00239617
COSM3983748	*NRG1*	Substitution - Missense	TSG, fusion	5	0	9.98E-04

Table 19 (continued).

COSM5776518	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3133345	*SLC34A2*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM85140	*RSPO2*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM3786996	*ARHGEF12*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM4950062	*EXT2*	Substitution - Missense	TSG	2	0	3.99E-04
COSM5008803	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3398773	*KMT2D*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM5694461	*DNM2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5575916	*ATRX*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5020013	*TET2*	Substitution - Missense	TSG	413	23	0.0916534
COSM5989864	*PTCH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4984943	*ERCC5*	Substitution - Missense	TSG	86	4	0.01877
COSM1026894	*PTPRT*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3365254	*BAP1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM145832	*PTPRB*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6023987	*CDH1*	Substitution - Missense	TSG	2	0	3.99E-04
COSM12507	*RP11-145E5.5*	Substitution - Missense	TSG	1	0	2.00E-04
COSM898730	*SPEN*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4989562	*PTPRT*	Substitution - Missense	TSG	2	0	3.99E-04
COSM3235559	*ATP2B3*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4117131	*MLH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1362025	*KMT2D*	Substitution - Missense	oncogene, TSG	41	1	0.00858626
COSM3766291	*TCF3*	Substitution - Missense	oncogene, TSG, fusion	42	0	0.00838658
COSM1391233	*TCF3*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM5008349	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2716441	*ERBB4*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM5903691	*WRN*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5945045	*TET2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2111018	*DDX10*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM2923605	*FANCA*	Substitution - Missense	TSG	1	0	2.00E-04
COSM28777	*KDM6A*	Substitution - Nonsense	oncogene, TSG	0	0	
COSM4718596	*PTCH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM907737	*ARID1A*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM1007064	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM202922	*NCOR2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3133335	*SLC34A2*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM5612735	*POLE*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6405476	*AXIN1*	Substitution - Missense	TSG	3	0	5.99E-04
COSM6200068	*CAMTA1*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM1213728	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4122123	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04

Table 20 (continued).

COSM4912352	*FAT1*	Substitution - Missense	TSG	2	0	3.99E-04
COSM4851464	*BCL9L*	Substitution - Missense	oncogene, TSG	6	0	0.00119808
COSM5019549	*BLM*	Substitution - Missense	TSG	237	10	0.0513179
COSM5981146	*NCOR1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1750941	*PAFAH1B3*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM5995709	*PTPN13*	Substitution - Missense	TSG	2	0	3.99E-04
COSM5869076	*FOXO3*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM242616	*RECQL4*	Substitution - Missense	oncogene, TSG	2	0	3.99E-04
COSM3952512	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	2	0	3.99E-04
COSM5415553	*PTPRK*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM1685386	*SH2B3*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3567190	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3899655	*WRN*	Substitution - Missense	TSG	3	0	5.99E-04
COSM914244	*SUFU*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2922693	*CBFA2T3*	Substitution - Missense	TSG, fusion	3	0	5.99E-04
COSM899738	*SDHB*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4588244	*NBN*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5981750	*PTK6*	Substitution - Missense	oncogene, TSG	6	0	0.00119808
COSM97172	*PRDM1*	Substitution - Missense	TSG	2	0	3.99E-04
COSM968803	*PALB2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6467686	*PTPRT*	Substitution - Missense	TSG	2	0	3.99E-04
COSM1235622	*MYH9*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM6455249	*WRN*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM24635	*ATM*	Substitution - Missense	TSG	6	0	0.00119808
COSM3172222	*PTPRK*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM22499	*ATM*	Substitution - Missense	TSG	13	0	0.00259585
COSM3754971	*CDH11*	Substitution - Missense	TSG, fusion	738	175	0.217252
COSM5743825	*PER1*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM5019671	*RANBP2*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM219132	*CBL*	Substitution - Nonsense	oncogene, TSG, fusion	1	0	2.00E-04
COSM943175	*PTPRB*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1736534	*PML*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM3235547	*ATP2B3*	Substitution - Missense	TSG	0	0	0
COSM4986833	*ERCC4*	Substitution - Missense	TSG	131	6	0.0285543
COSM5784452	*ZFHX3*	Substitution - Missense	TSG	1	0	2.00E-04
COSM1579580	*FANCA*	Substitution - Missense	TSG	109	4	0.0233626
COSM982272	*BRIP1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM232262	*ERBB4*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM3663500	*EIF3E*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM4582279	*CHEK2*	Substitution - Missense	TSG	2	0	3.99E-04
COSM3304542	*KMT2C*	Substitution - Missense	TSG	2	0	3.99E-04

Table 21 (continued).

COSM3466980	*ERCC5*	Substitution - Missense	TSG	5	0	9.98E-04
COSM3753566	*ERCC5*	Substitution - Missense	TSG	1118	346	0.361422
COSM5587819	*FES*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM1049765	*TET2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5537808	*PTPRB*	Substitution - Missense	TSG	1	0	2.00E-04
COSM17485	*PTCH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3909209	*PTCH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4605861	*NOTCH2*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM163645	*POT1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3772152	*TSC2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM48575	*MSH6*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4606482	*KMT2C*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4169988	*NOTCH1*	Substitution - Missense	oncogene, TSG, fusion	1	0	2.00E-04
COSM282501	*LRP1B*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4631416	*PTCH1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM21301	*ATM*	Substitution - Missense	TSG	2	0	3.99E-04
COSM977450	*NF1*	Substitution - Missense	TSG, fusion	1	0	2.00E-04
COSM430330	*SH2B3*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4987639	*ERCC4*	Substitution - Missense	TSG	29	0	0.00579073
COSM1131275	*FAT4*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5019066	*TSC2*	Substitution - Missense	TSG	4	0	7.99E-04
COSM1222520	*PTPN13*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3639952	*PMS2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM3457812	*NCOR2*	Substitution - Missense	TSG	1	0	2.00E-04
COSM2716452	*ERBB4*	Substitution - Missense	oncogene, TSG	1	0	2.00E-04
COSM1979240	*EXT2*	Substitution - Missense	TSG	4	0	7.99E-04
COSM249142	*DNMT3A*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5005948	*ERCC4*	Substitution - Missense	TSG	4	0	7.99E-04
COSM3756940	*CBLC*	Substitution - Missense	oncogene, TSG	449	24	0.0992412
COSM5627064	*BLM*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6376158	*KMT2C*	Substitution - Missense	TSG	2	0	3.99E-04
COSM922745	*ATM*	Substitution - Missense	TSG	1	0	2.00E-04
COSM4606553	*SH2B3*	Substitution - Missense	TSG	6	0	0.00119808
COSM3399055	*PTPRB*	Substitution - Missense	TSG	1	0	2.00E-04
COSM5569099	*ZFHX3*	Substitution - Missense	TSG	0	0	0
COSM5016779	*BRIP1*	Substitution - Missense	TSG	1	0	2.00E-04
COSM6357850	*LZTR1*	Substitution - Nonsense	TSG	1	0	2.00E-04
COSM3149630	*MYH9*	Substitution - Missense	TSG, fusion	2	0	3.99E-04
COSM1583141	*DNMT3A*	Substitution - Missense	TSG	1	0	2.00E-04

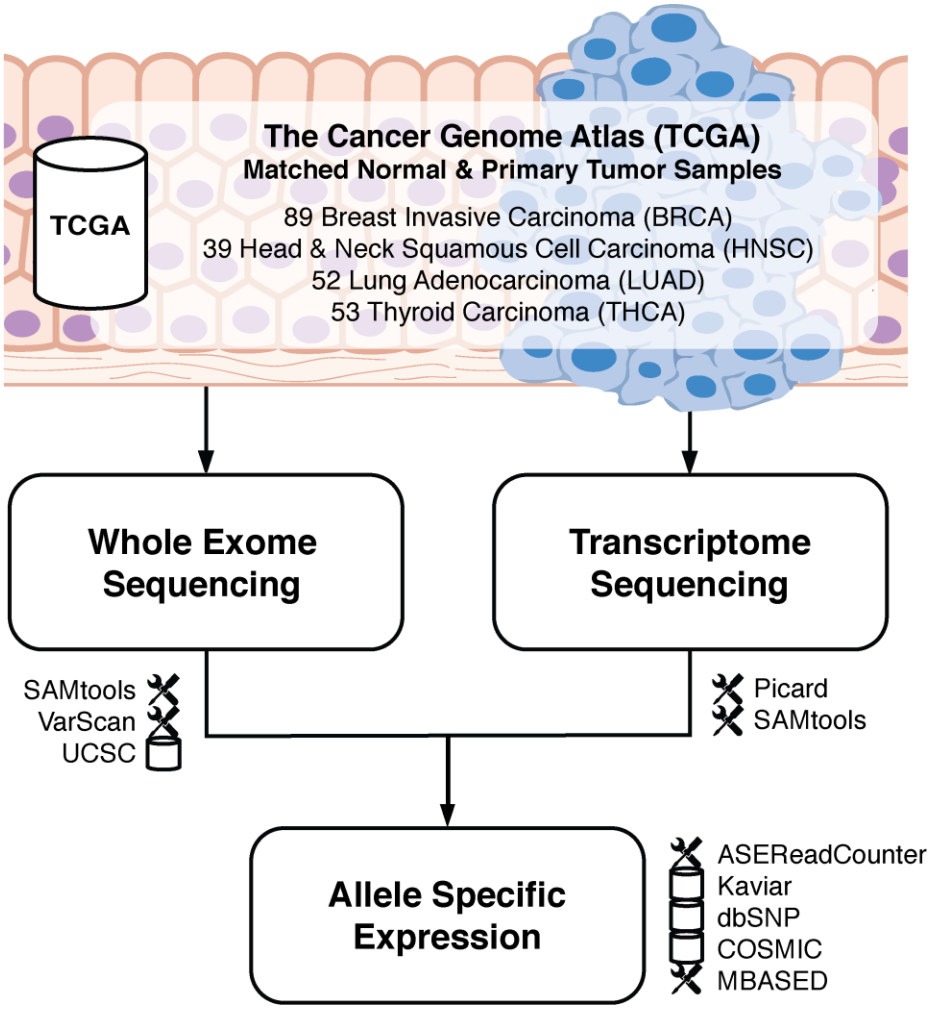

Figure 33 ASE workflow used in this study.

Matched normal and primary tumor samples for four cancer types were analyzed for allele-specific expression using whole exome (DNA-seq) and transcriptome (RNA-seq) data. DNA-seq data was used to identify heterozygous sites in the exome and RNA-seq data was analyzed to compare expression of reference vs alternative alleles at those sites. The core bioinformatics databases and tools used for each stage of the pipeline are indicated.

Table 22 Percent of genes displaying ASE in 233 TCGA patients.

	Pattern	% Total Genes	% All COSMIC	% TSG	% Oncogene	% Fusion
BRCA	1	9.6	9.3	9.1	8.3	9.3
	2	0.5	0.5	0.4	0.3	0.7
	3	0.0	0.0	0.0	0.0	0.0
	4	3.4	3.3	2.8	3.4	3.7
	5	2.7	2.2	2.1	2.5	2.1
	6	0.1	0.1	0.0	0.0	0.1
	No ASE	83.7	84.6	85.6	85.5	84.0
HNSC	1	12.4	12.5	11.1	13.6	13.1
	2	0.4	0.3	0.1	0.6	0.6
	3	0.1	0.1	0.1	0.0	0.2
	4	4.2	4.1	4.5	3.9	3.8
	5	2.6	2.0	2.0	2.1	1.7
	6	0.1	0.2	0.1	0.1	0.3
	No ASE	80.3	80.9	82.1	79.7	80.3
LUAD	1	13.4	13.1	13.2	12.5	14.0
	2	0.3	0.3	0.1	0.4	0.4
	3	0.0	0.1	0.0	0.0	0.1
	4	5.2	5.7	5.6	5.5	5.8
	5	1.7	1.3	1.4	1.6	1.4
	6	0.1	0.0	0.0	0.1	0.0
	No ASE	79.3	79.5	79.6	80.0	78.3
THCA	1	3.3	2.9	2.8	3.3	3.8
	2	0.3	0.3	0.1	0.2	0.6
	3	0.0	0.0	0.0	0.0	0.0
	4	0.5	0.5	0.2	0.5	1.0
	5	2.6	2.2	2.6	2.3	2.2
	6	0.0	0.0	0.0	0.0	0.0
	No ASE	93.3	94.1	94.3	93.8	92.4

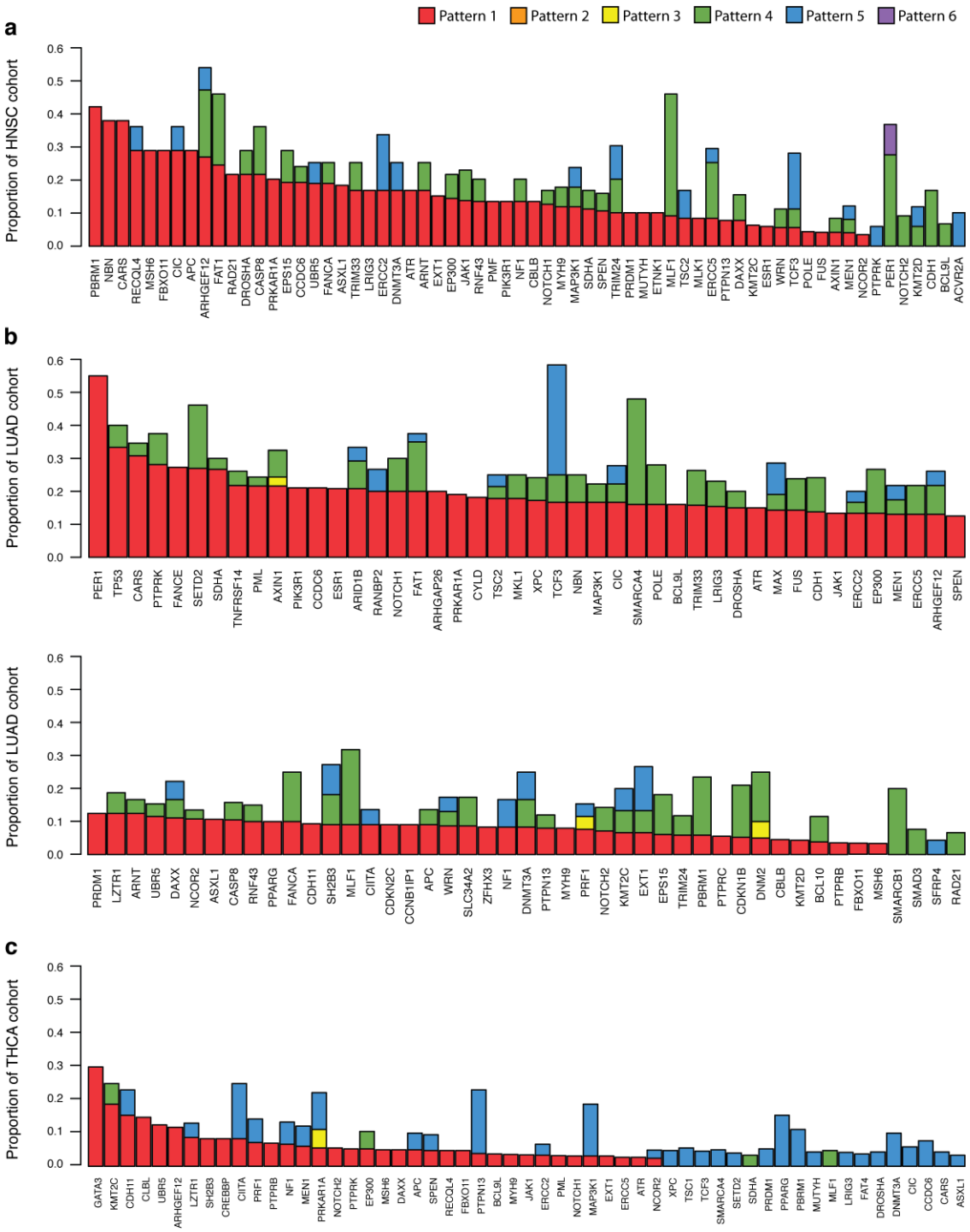

Figure 34 Frequency of ASE in tumor suppressor genes.

Gene level ASE was computed as described in the Materials and Methods section. a, The proportion of head and neck cancer patients with ASE in 64 TSGs b, The proportion of lung cancer patients with ASE in 89 TSGs c, The proportion of thyroid cancer patients with ASE in 55 TSGs.

Table 23 DNA-sequencing (DNA-seq) and RNA-sequencing (RNA-seq) data sources for the nine TCGA patients analyzed for mechanism of ASE in this study.

ID	TCGA Barcode	Cancer Type	Sex	Age	Race	Sample Type[a]	Seq Depth[b]	Read Len.[c]
Breast 1	TCGA-BH-A0B3-11B-21D-A128-09	Breast Invasive Carcinoma	F	53	White	NT-G	42.4	100
	TCGA-BH-A0B3-11B-21R-A089-07					NT-R	5.5	50
	TCGA-BH-A0B3-11B-21W-A100-09					NT-X	8.1	100
	TCGA-BH-A0B3-01A-11D-A128-09					TP-G	40.2	100
	TCGA-BH-A0B3-01B-21R-A089-07					TP-R	5.4	50
	TCGA-BH-A0B3-01A-11W-A071-09					TP-X	10.5	100
Breast 2	TCGA-BH-A0BW-11A-12D-A314-09		F	71	Black	NT-G	54.1	100
	TCGA-BH-A0BW-11A-12R-A115-07					NT-R	7	50
	TCGA-BH-A0BW-11A-12D-A10Y-09					NT-X	14.7	100
	TCGA-BH-A0BW-01A-11D-A10Y-09					TP-G	46.1	100
	TCGA-BH-A0BW-01A-12R-A115-07					TP-R	7.3	50
	TCGA-BH-A0BW-01A-11D-A10Y-09					TP-X	15.9	100
Breast 3	TCGA-BH-A0DT-11A-12D-A12B-09		F	41	White	NT-G	63.3	100
	TCGA-BH-A0DT-11A-12R-A12D-07					NT-R	7.7	50
	TCGA-BH-A0DT-11A-12D-A12B-09					NT-X	22.7	100
	TCGA-BH-A0DT-01A-21D-A12B-09					TP-G	79.9	100
	TCGA-BH-A0DT-01A-21R-A12D-07					TP-R	6.6	50
	TCGA-BH-A0DT-01A-21D-A12B-09					TP-X	21.7	100
Head 1	TCGA-CV-7255-11A-01D-2276-10	Head and Neck Squamous Cell Carcinoma	F	32	White	NT-G	6.9	101
	TCGA-CV-7255-11A-01R-2016-07					NT-R	7.5	48
	TCGA-CV-7255-11A-01D-2012-08					NT-X	27.3	76
	TCGA-CV-7255-01A-11D-2276-10					TP-G	5.8	101
	TCGA-CV-7255-01A-11R-2016-07					TP-R	7.1	48
	TCGA-CV-7255-01A-11D-2012-08					TP-X	28.9	76
Head 2	TCGA-CV-7416-11A-01D-2334-08		F	29	White	NT-G	7.7	101
	TCGA-CV-7416-11A-01R-2081-07					NT-R	5.9	48
	TCGA-CV-7416-11A-01D-2078-08					NT-X	23.9	76
	TCGA-CV-7416-01A-11D-2334-08					TP-G	28.6	101
	TCGA-CV-7416-01A-11R-2081-07					TP-R	6	48
	TCGA-CV-7416-01A-11D-2078-08					TP-X	25.0	76
Head 3	TCGA-CV-6959-11A-01D-1911-02		M	48	White	NT-G	38.3	51
	TCGA-CV-6959-11A-01R-1915-07					NT-R	8.5	48
	TCGA-CV-6959-11A-01D-1912-08					NT-X	26.8	76
	TCGA-CV-6959-01A-11D-1911-02					TP-G	31.4	51
	TCGA-CV-6959-01A-11R-1915-07					TP-R	6.6	48
	TCGA-CV-6959-01A-11D-1912-08					TP-X	28.0	76
Lung 1	TCGA-44-6776-11A-01D-1853-02	Lung Adenocarcinoma	F	60	White	NT-G	38.9	51
	TCGA-44-6776-11A-01R-1858-07					NT-R	5.4	48
	TCGA-44-6776-11A-01D-1855-08					NT-X	29.1	76
	TCGA-44-6776-01A-11D-1853-02					TP-G	6.9	51

Table 243 (continued).

	TCGA-44-6776-01A-11R-1858-07					TP-R	7.4	48
	TCGA-44-6776-01A-11D-1855-08					TP-X	30.8	76
Lung 2	TCGA-50-5932-11A-01D-1753-08		M	75	White	NT-G	34.6	101
	TCGA-50-5932-11A-01R-1755-07					NT-R	4.2	48
	TCGA-50-5932-11A-01D-1753-08					NT-X	29.3	76
	TCGA-50-5932-01A-11D-1753-08					TP-G	44.5	101
	TCGA-50-5932-01A-11R-1755-07					TP-R	7.4	48
	TCGA-50-5932-01A-11D-1753-08					TP-X	33.6	76
Lung 3	TCGA-55-6984-11A-01D-1945-08		F	NA	White	NT-G	36.2	101
	TCGA-55-6984-11A-01R-1949-07					NT-R	4.9	48
	TCGA-55-6984-11A-01D-1945-08					NT-X	23.1	76
	TCGA-55-6984-01A-11D-1945-08					TP-G	41	101
	TCGA-55-6984-01A-11R-1949-07					TP-R	5.2	48
	TCGA-55-6984-01A-11D-1945-08					TP-X	17.8	76

[a]NT-G=Normal tissue WGS, NT-R=Normal tissue RNA-seq, NT-X=Normal tissue WXS, TP-G=Tumor primary WGS, TP-R=Tumor primary RNA-seq, TP-X=Tumor primary WXS; [b]Sequencing depth a.k.a coverage; [c]Read length measured in base pairs

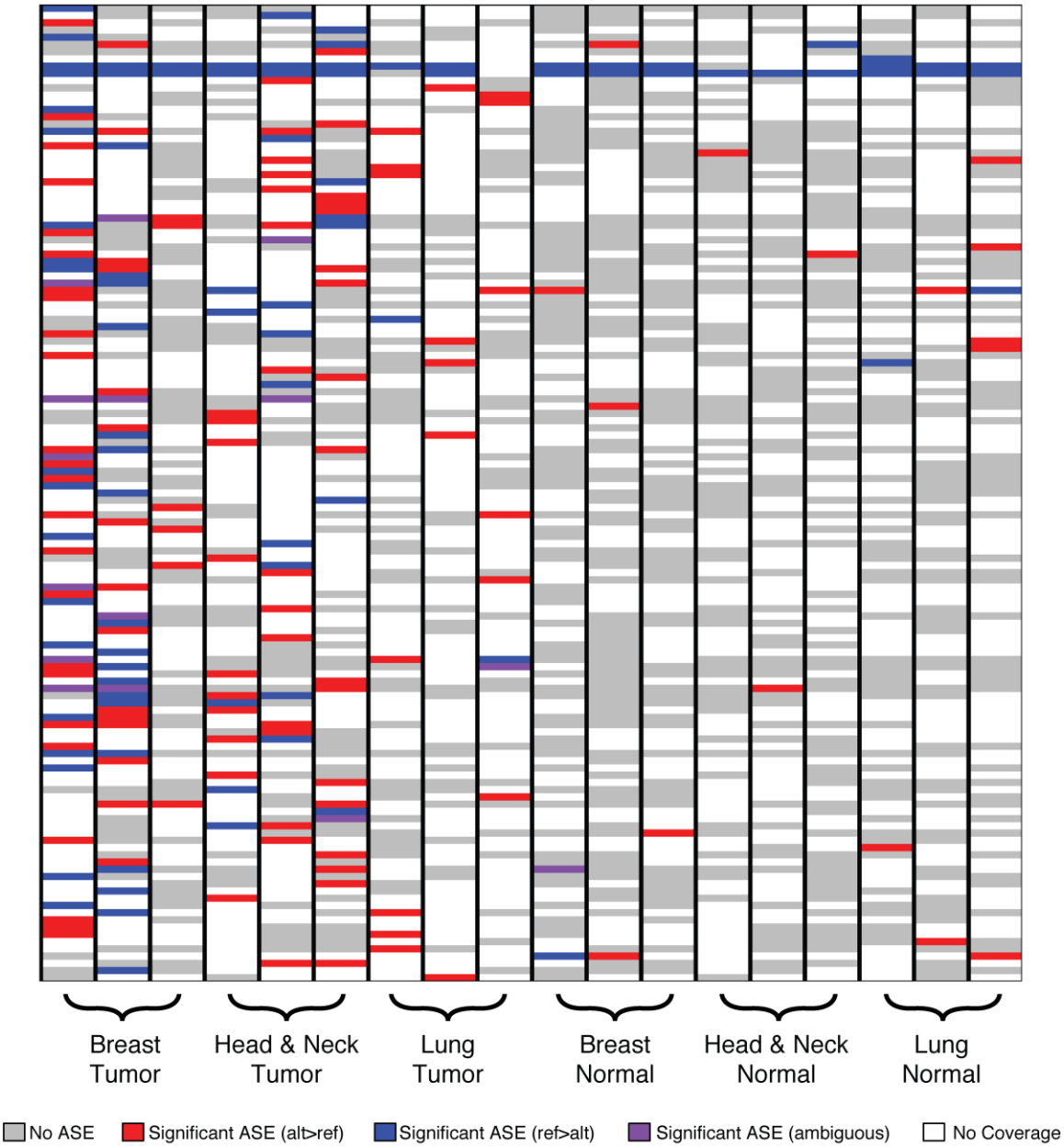

Figure 35 Heatmap of COSMIC genes across 9 patients analyzed for mechanism of ASE.

Rows are genes and columns are samples. Genes are sorted by genomic location. Samples are grouped by tissue type, tumor (left) and normal (right) and by cancer (lung, breast, head & neck). Significant ASE (P < 0.05) genes are shown in red (alt > ref) and blue (ref > alt). Note some genes have significant ASE however their direction is ambiguous (purple) because of an equal number of reference and alternative SNPs on each pseudo-haplotype.

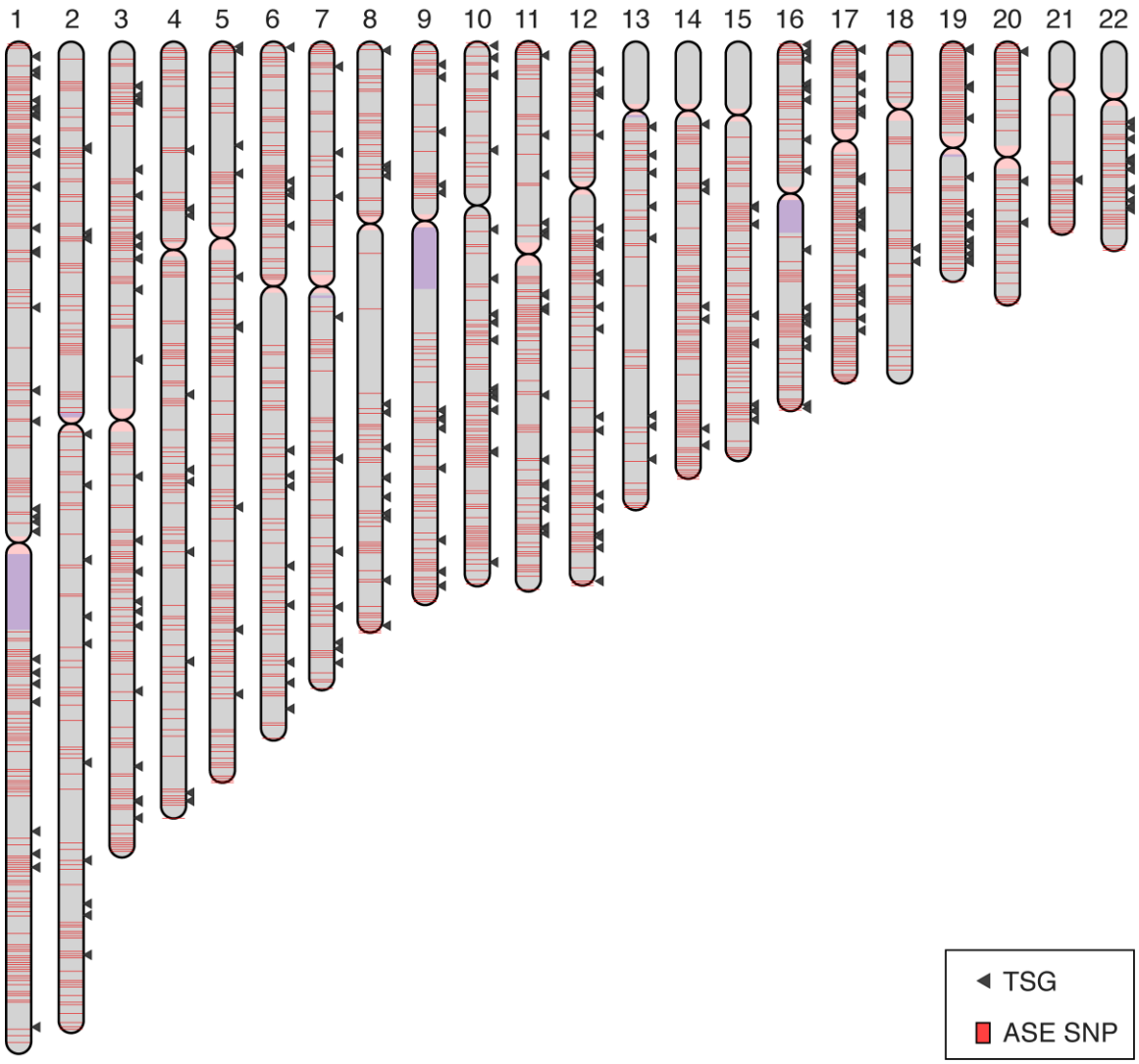

Figure 36 Distribution of ASE SNPs in the genome from 9 patients studied for mechanism.

Genome Ideogram showing locations of ASE SNPs marked in red; TSGs denoted with triangle. Chromosomes are rendered with pink centromeres and purple heterochromatin bands.

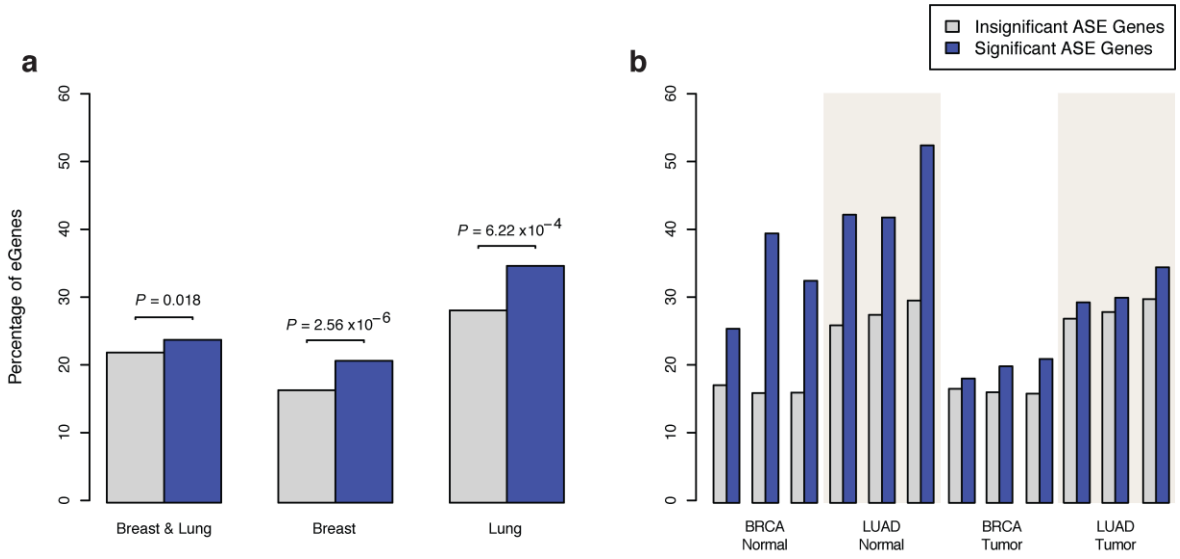

Figure 37 ASE eGenes.

Heterozygous eQTLs associated with protein coding genes (eGenes) were identified in patient samples using whole genome sequencing (WGS). a, Percentages of ASE genes associated with eQTLs in breast and lung cancer patient samples combined. b, Percentages of ASE genes associated with eQTLs broken down by individual patient sample.

Table 25 ASE Genes possibly explained by eQTLs or differential methylation of CpG Islands.

Patient	Total ASE Genes	ASE Genes w phased eQTL genotypes that explain Pattern	Percentage of ASE putatively explained by eQTLs	ASE Genes w beta value | fold change | > 1.3	Percentage of ASE correlated
Breast 1	845	8	0.9%	113	13.4
Breast 2	834	6	0.7%	25	3.0
Breast 3	296	5	1.7%	9	3.4
Head & Neck 1	447	*eQTLs not available*		87	19.5
Head & Neck 2	NA	*eQTLs not available*		*Methylation data not available*	
Head & Neck 3	540	*eQTLs not available*		65	12.0
Lung 1	458	25	5.4%	*Methylation data not available*	
Lung 2	131	1	5.1%	15	13.0
Lung 3	201	4	5.4%	*Methylation data not available*	
Total	3,752	49	**1.8%**	317	**10.2**

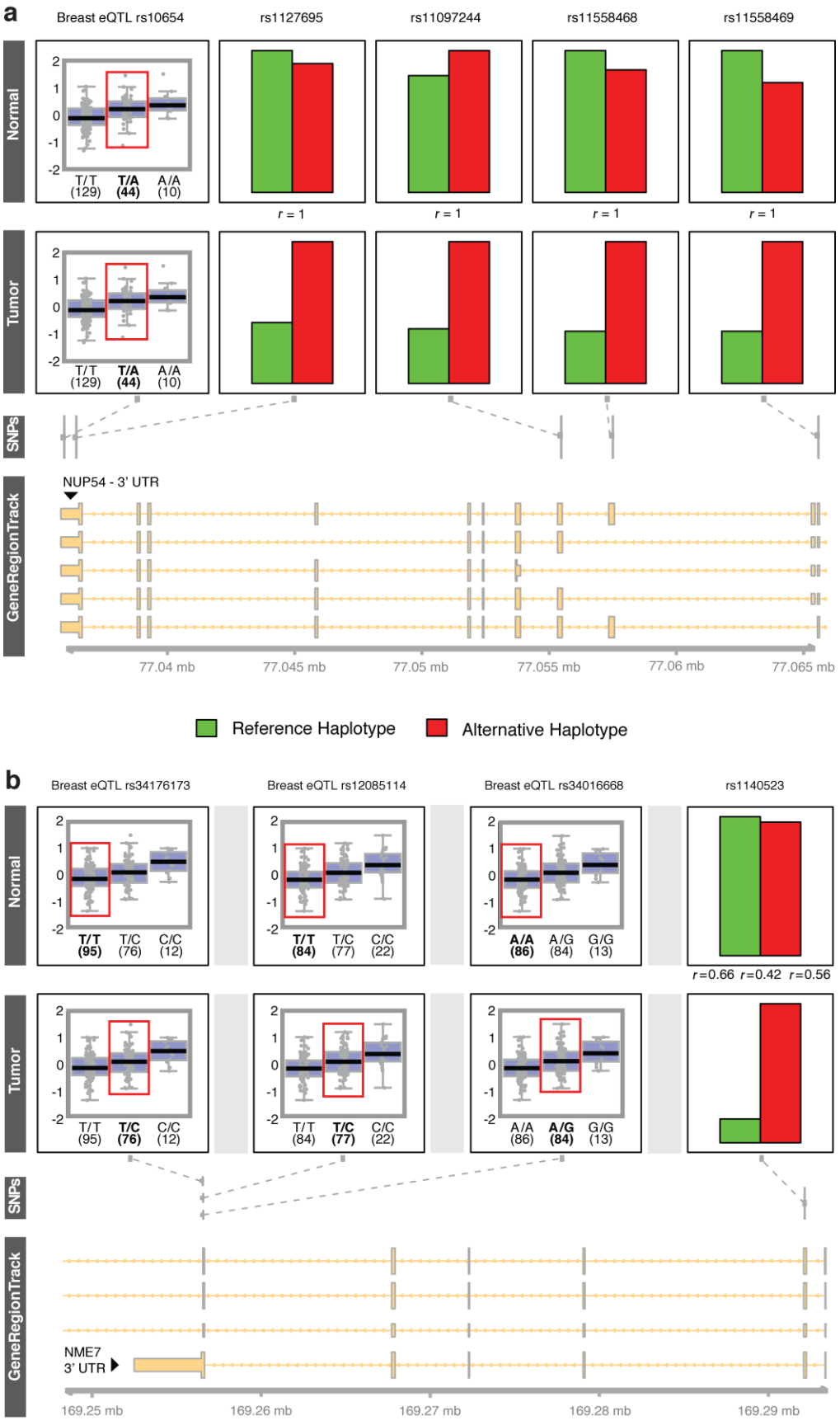
a
Breast eQTL rs10654
rs1127695
rs11097244
rs11558468
rs11558469
Normal
Tumor
T/T (129)
T/A (44)
A/A (10)
r = 1
SNPs
GeneRegionTrack
NUP54 - 3' UTR
77.04 mb
77.045 mb
77.05 mb
77.055 mb
77.06 mb
77.065 mb
Reference Haplotype
Alternative Haplotype
b
Breast eQTL rs34176173
Breast eQTL rs12085114
Breast eQTL rs34016668
rs1140523
T/T (95)
T/C (76)
C/C (12)
T/T (84)
T/C (77)
C/C (22)
A/A (86)
A/G (84)
G/G (13)
r=0.66 r=0.42 r=0.56
A/A (86)
A/G (84)
G/G (13)
NME7
3' UTR
169.25 mb
169.26 mb
169.27 mb
169.28 mb
169.29 mb

Figure 38 eQTLs in cis and trans with ASE genes.

Specific genomic regions spanning eQTLs and ASE SNPs were plotted using AllelicImbalance as described in the Materials and Methods. The top panels contain GTEx single tissue eQTL box plots and the relative expression of alleles for linked SNPs. The grey lines in the middle panel show where the SNP locations lie in relation to the appropriate genome track shown beneath in yellow. a, A heterozygous eQTL present in the 3' UTR of NUP54 in both the normal and tumor samples of a breast invasive carcinoma patient. b, Three eQTLs present downstream from NME7 which are homozygous reference in the normal sample and heterozygous in the tumor sample of a breast invasive carcinoma patient.

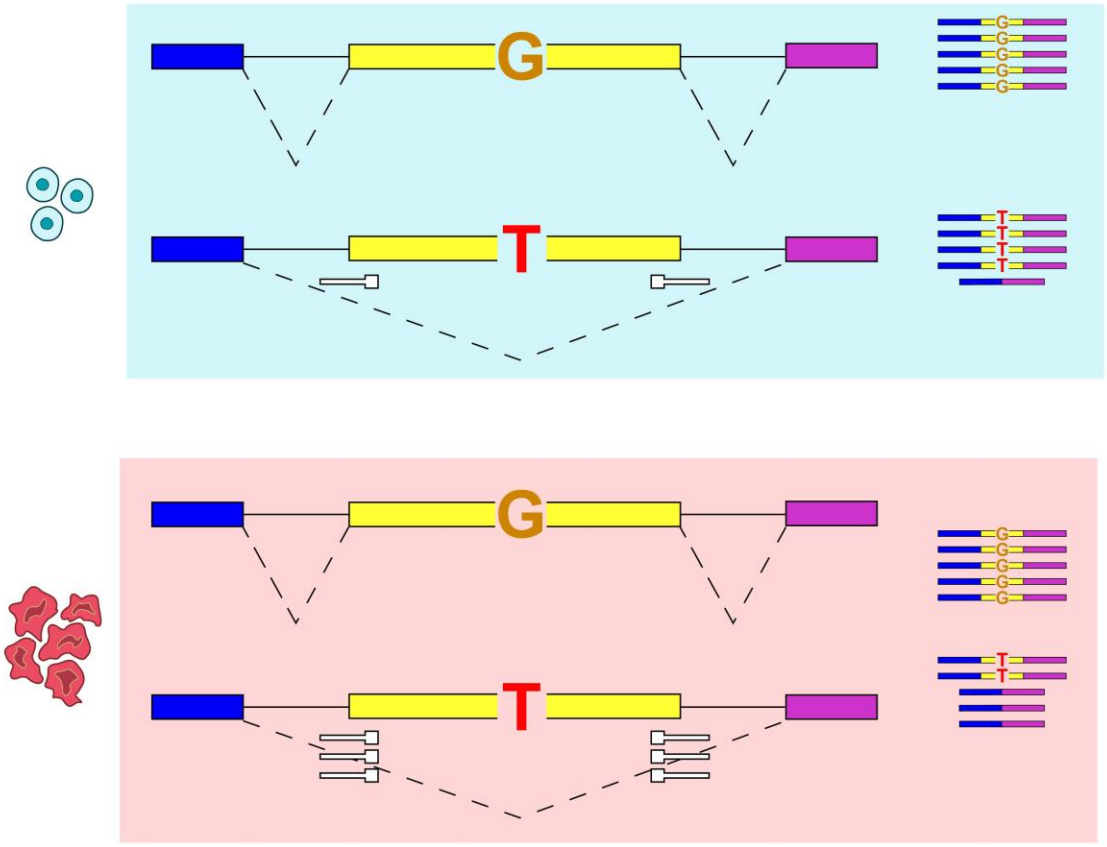

Figure 39 Model for antisense induced allele-specific exon skipping and its contribution to ASE.

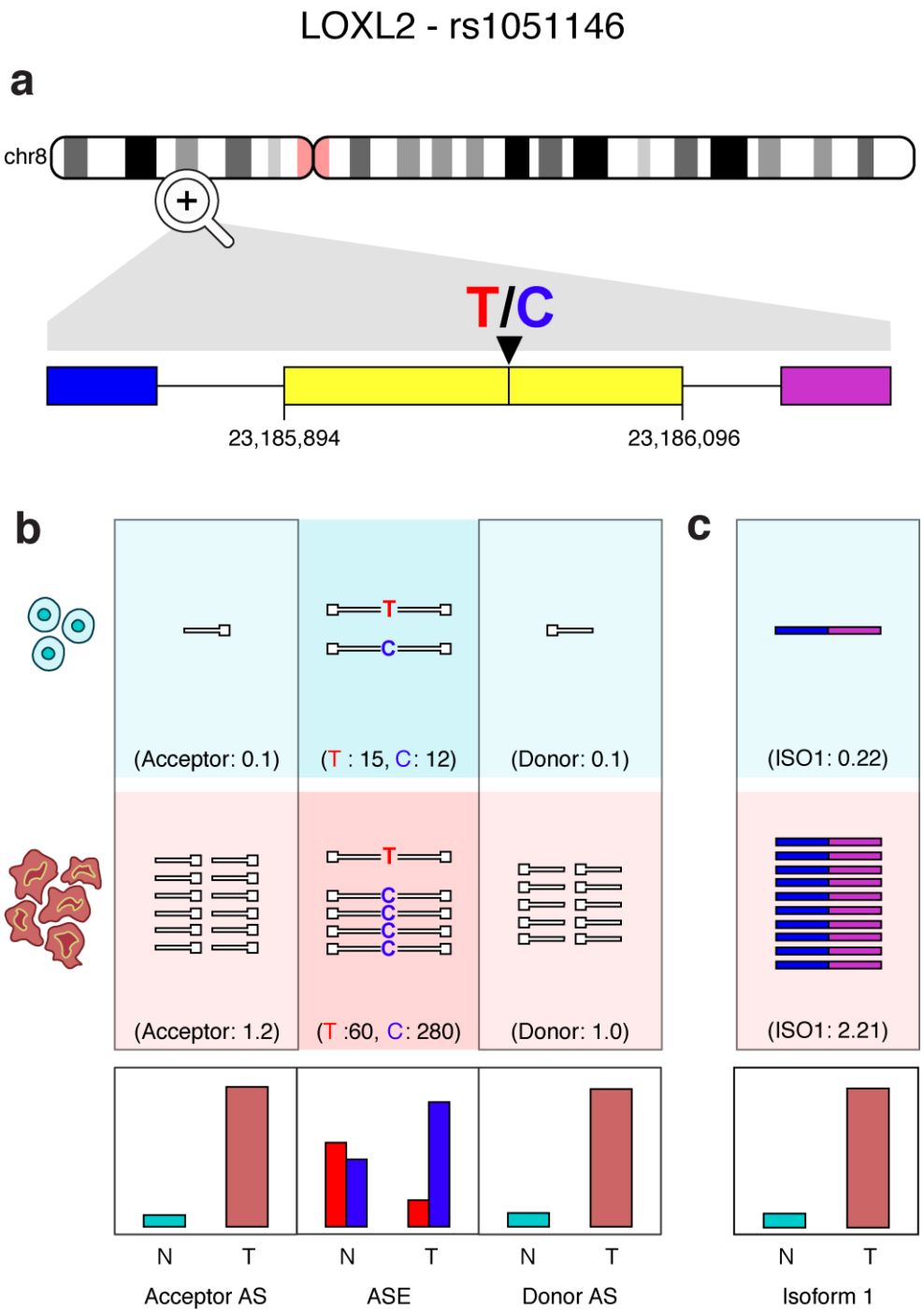

Figure 40 *LOXL2* exon skipping correlates with ASE in a breast adenocarcinoma patient.

a, An exon skipping event in exon 6 of TNC in a breast cancer patient (TCGA-BH-A0B3). b, Antisense reads mapping to donor and acceptor sites are quantified, alongside the ASE

locus within the exon. c, Quantification of reads supporting the isoform missing exon 6. Relative expression plots are shown for antisense RNA, ASE and isoforms below.

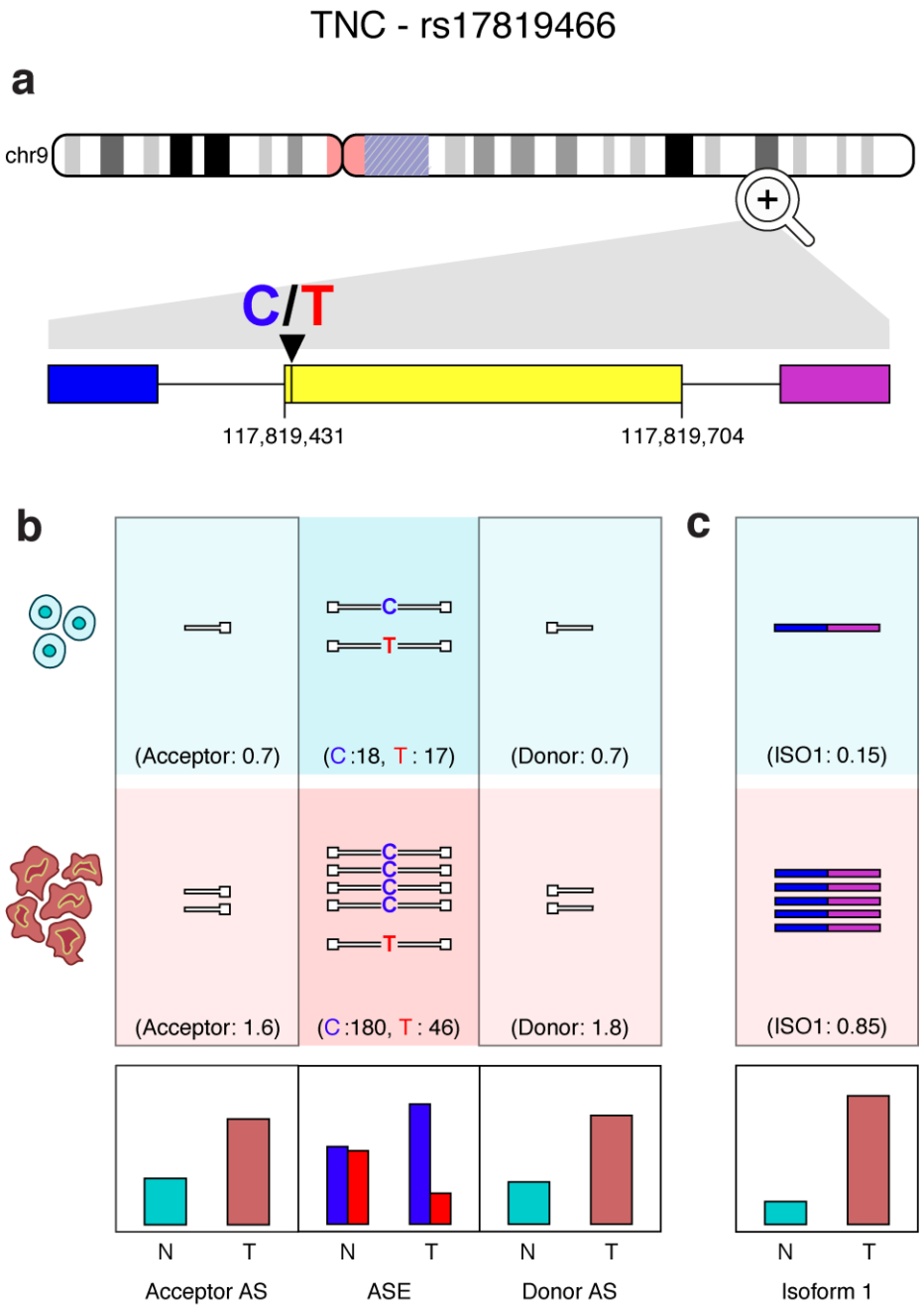

Figure 41 *TNC* exon skipping correlates with ASE in a breast adenocarcinoma patient.

a, An exon skipping event in exon 15 of TNC in a breast cancer patient (TCGA-BH-A0DT). b, Antisense reads mapping to donor and acceptor sites are quantified, alongside the ASE locus within the exon. c, Quantification of reads supporting the isoform missing exon 15. Relative expression plots are shown for antisense RNA, ASE and isoforms below.

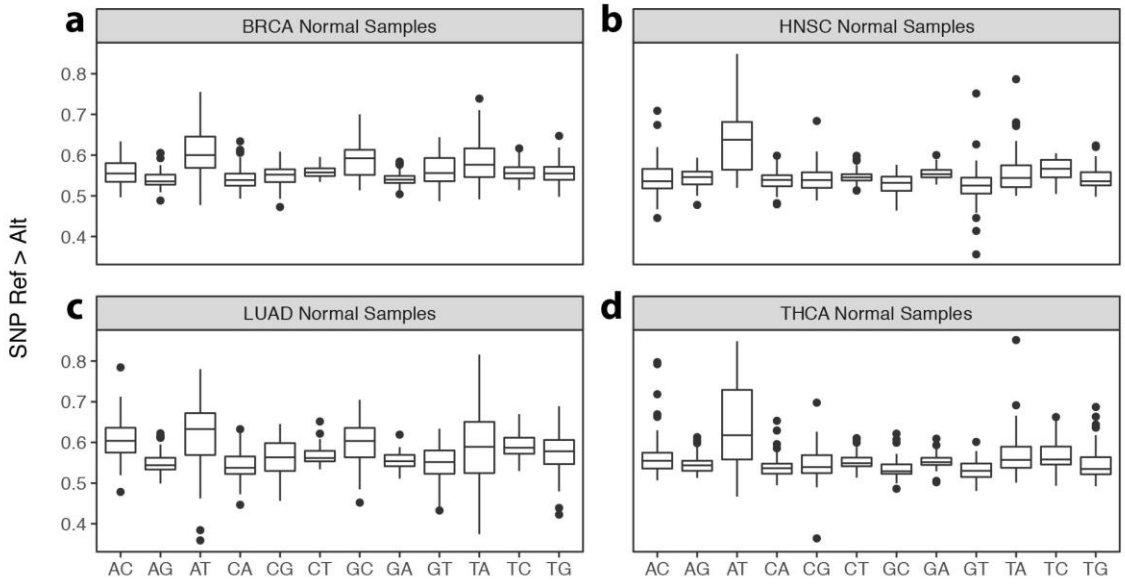

Figure 42 Allelic ratios for all possible nucleotide combinations.

Distribution of proportion of sites where reference allele count is greater than alternative allele count in normal samples for a, breast cancer patients. b, head and neck cancer patients. c, lung cancer patients. d, thyroid cancer patients.

Table 26 Metadata for 233 TCGA patients analyzed in this study.

TCGA Barcode	Cancer Type	Sex	Age	Race
TCGA-A7-A0CE	BRCA	female	57	white
TCGA-A7-A0D9	BRCA	female	37	white
TCGA-A7-A0DB	BRCA	female	56	white
TCGA-A7-A13E	BRCA	female	62	white
TCGA-A7-A13F	BRCA	female	44	white
TCGA-A7-A13G	BRCA	female	79	white
TCGA-AC-A23H	BRCA	female	90	white
TCGA-AC-A2FB	BRCA	female	65	white
TCGA-AC-A2FF	BRCA	female	NA	asian
TCGA-AC-A2FM	BRCA	female	87	white
TCGA-BH-A0AU	BRCA	female	45	white
TCGA-BH-A0AY	BRCA	female	62	white
TCGA-BH-A0AZ	BRCA	female	47	white
TCGA-BH-A0B3	BRCA	female	53	white
TCGA-BH-A0B5	BRCA	female	40	white
TCGA-BH-A0B7	BRCA	female	42	white
TCGA-BH-A0B8	BRCA	female	64	white
TCGA-BH-A0BA	BRCA	female	51	white

Table 15 (continued).

TCGA-BH-A0BC	BRCA	female	60	white
TCGA-BH-A0BJ	BRCA	female	41	white
TCGA-BH-A0BM	BRCA	female	54	white
TCGA-BH-A0BQ	BRCA	female	39	white
TCGA-BH-A0BT	BRCA	female	56	white
TCGA-BH-A0BV	BRCA	female	78	white
TCGA-BH-A0BW	BRCA	female	71	black or african american
TCGA-BH-A0BZ	BRCA	female	59	white
TCGA-BH-A0C0	BRCA	female	63	white
TCGA-BH-A0C3	BRCA	female	47	white
TCGA-BH-A0DD	BRCA	male	58	white
TCGA-BH-A0DG	BRCA	female	30	black or african american
TCGA-BH-A0DH	BRCA	female	64	white
TCGA-BH-A0DK	BRCA	female	49	white
TCGA-BH-A0DL	BRCA	female	64	white
TCGA-BH-A0DP	BRCA	female	60	white
TCGA-BH-A0DQ	BRCA	female	42	white
TCGA-BH-A0DT	BRCA	female	41	white
TCGA-BH-A0DV	BRCA	female	54	white
TCGA-BH-A0DZ	BRCA	female	43	white
TCGA-BH-A0E0	BRCA	female	38	white
TCGA-BH-A0E1	BRCA	female	52	white
TCGA-BH-A0H5	BRCA	female	45	white
TCGA-BH-A0H7	BRCA	female	65	white
TCGA-BH-A0H9	BRCA	female	69	white
TCGA-BH-A0HA	BRCA	female	31	white
TCGA-BH-A0HK	BRCA	female	81	white
TCGA-BH-A18J	BRCA	female	56	white
TCGA-BH-A18K	BRCA	female	46	white
TCGA-BH-A18L	BRCA	female	50	white
TCGA-BH-A18M	BRCA	female	39	white
TCGA-BH-A18N	BRCA	female	88	white
TCGA-BH-A18P	BRCA	female	60	white
TCGA-BH-A18Q	BRCA	female	56	white
TCGA-BH-A18S	BRCA	female	79	white
TCGA-BH-A18U	BRCA	female	72	white
TCGA-BH-A18V	BRCA	female	48	white
TCGA-BH-A1EN	BRCA	female	78	white
TCGA-BH-A1EO	BRCA	female	68	white
TCGA-BH-A1ET	BRCA	female	55	white
TCGA-BH-A1EV	BRCA	female	45	white

Table 15 (continued).

TCGA-BH-A1EW	BRCA	female	38	white
TCGA-BH-A1F0	BRCA	female	80	white
TCGA-BH-A1F2	BRCA	female	53	white
TCGA-BH-A1F6	BRCA	female	51	white
TCGA-BH-A1F8	BRCA	female	90	white
TCGA-BH-A1FC	BRCA	female	78	black or african american
TCGA-BH-A1FD	BRCA	female	68	white
TCGA-BH-A1FE	BRCA	female	31	white
TCGA-BH-A1FG	BRCA	female	88	white
TCGA-BH-A1FH	BRCA	female	47	white
TCGA-BH-A1FJ	BRCA	female	66	white
TCGA-BH-A1FM	BRCA	female	44	white
TCGA-BH-A1FN	BRCA	female	34	white
TCGA-BH-A1FR	BRCA	female	73	white
TCGA-BH-A1FU	BRCA	female	44	white
TCGA-BH-A203	BRCA	female	78	white
TCGA-BH-A204	BRCA	female	80	white
TCGA-BH-A208	BRCA	female	48	white
TCGA-BH-A209	BRCA	female	77	black or african american
TCGA-E2-A153	BRCA	female	51	white
TCGA-E2-A158	BRCA	female	43	white
TCGA-E2-A15I	BRCA	female	44	white
TCGA-E2-A15K	BRCA	female	58	white
TCGA-E2-A15M	BRCA	female	66	white
TCGA-E2-A1BC	BRCA	female	63	not reported
TCGA-E2-A1LB	BRCA	female	41	black or african american
TCGA-E2-A1LH	BRCA	female	59	white
TCGA-E2-A1LS	BRCA	female	46	white
TCGA-GI-A2C8	BRCA	female	63	white
TCGA-GI-A2C9	BRCA	female	58	black or african american
TCGA-CV-6933	HNSC	male	53	white
TCGA-CV-6934	HNSC	female	66	white
TCGA-CV-6935	HNSC	male	67	white
TCGA-CV-6936	HNSC	male	68	white
TCGA-CV-6938	HNSC	male	87	white
TCGA-CV-6939	HNSC	male	60	white
TCGA-CV-6943	HNSC	male	74	white
TCGA-CV-6955	HNSC	female	74	white
TCGA-CV-6956	HNSC	male	67	white
TCGA-CV-6959	HNSC	male	48	white
TCGA-CV-6960	HNSC	male	49	black or african american

Table 15 (continued).

TCGA-CV-6961	HNSC	male	61	white
TCGA-CV-6962	HNSC	male	66	white
TCGA-CV-7091	HNSC	male	54	white
TCGA-CV-7097	HNSC	male	53	white
TCGA-CV-7101	HNSC	male	80	white
TCGA-CV-7103	HNSC	male	49	white
TCGA-CV-7177	HNSC	female	82	white
TCGA-CV-7178	HNSC	female	64	white
TCGA-CV-7183	HNSC	male	53	white
TCGA-CV-7235	HNSC	male	67	white
TCGA-CV-7238	HNSC	female	69	white
TCGA-CV-7242	HNSC	female	60	white
TCGA-CV-7245	HNSC	male	62	white
TCGA-CV-7250	HNSC	male	64	white
TCGA-CV-7252	HNSC	female	62	black or african american
TCGA-CV-7255	HNSC	female	32	white
TCGA-CV-7261	HNSC	male	57	white
TCGA-CV-7406	HNSC	male	50	white
TCGA-CV-7416	HNSC	female	29	white
TCGA-CV-7423	HNSC	male	65	white
TCGA-CV-7424	HNSC	male	67	asian
TCGA-CV-7425	HNSC	female	77	white
TCGA-CV-7432	HNSC	male	79	white
TCGA-CV-7434	HNSC	male	64	white
TCGA-CV-7437	HNSC	male	77	not reported
TCGA-CV-7438	HNSC	female	87	white
TCGA-CV-7440	HNSC	male	38	white
TCGA-HD-8635	HNSC	female	61	white
TCGA-22-4593	LUAD	male	77	white
TCGA-22-4609	LUAD	male	81	white
TCGA-22-5471	LUAD	male	76	white
TCGA-22-5478	LUAD	male	79	not reported
TCGA-22-5481	LUAD	female	72	white
TCGA-22-5482	LUAD	male	81	white
TCGA-22-5483	LUAD	male	74	white
TCGA-22-5489	LUAD	male	64	white
TCGA-22-5491	LUAD	male	74	white
TCGA-33-4587	LUAD	female	63	white
TCGA-33-6737	LUAD	male	71	white
TCGA-38-4625	LUAD	female	66	white
TCGA-38-4626	LUAD	female	57	white

Table 15 (continued).

TCGA-38-4627	LUAD	female	64	white
TCGA-38-4632	LUAD	male	42	black or african american
TCGA-39-5040	LUAD	male	59	white
TCGA-43-6771	LUAD	male	85	white
TCGA-44-2655	LUAD	female	65	white
TCGA-44-2657	LUAD	female	74	white
TCGA-44-2662	LUAD	male	65	white
TCGA-44-5645	LUAD	female	61	black or african american
TCGA-44-6777	LUAD	female	85	white
TCGA-44-6778	LUAD	male	59	black or african american
TCGA-49-4490	LUAD	female	45	white
TCGA-49-6745	LUAD	male	82	white
TCGA-49-6761	LUAD	female	68	white
TCGA-50-5930	LUAD	male	47	white
TCGA-50-5931	LUAD	female	75	white
TCGA-50-5932	LUAD	male	75	white
TCGA-50-5933	LUAD	male	72	white
TCGA-50-5935	LUAD	female	86	white
TCGA-50-5936	LUAD	male	58	white
TCGA-50-5939	LUAD	male	85	white
TCGA-50-6595	LUAD	female	74	white
TCGA-51-4079	LUAD	female	73	black or african american
TCGA-55-6970	LUAD	female	67	white
TCGA-55-6972	LUAD	male	72	white
TCGA-55-6975	LUAD	male	NA	white
TCGA-55-6978	LUAD	male	NA	white
TCGA-55-6982	LUAD	female	NA	white
TCGA-55-6984	LUAD	female	NA	white
TCGA-55-6986	LUAD	female	NA	white
TCGA-56-7580	LUAD	male	84	white
TCGA-56-7582	LUAD	male	84	white
TCGA-77-7138	LUAD	male	67	not reported
TCGA-77-8007	LUAD	male	68	not reported
TCGA-85-7710	LUAD	female	59	white
TCGA-90-7767	LUAD	male	56	white
TCGA-91-6829	LUAD	male	79	white
TCGA-91-6831	LUAD	male	66	white
TCGA-91-6836	LUAD	female	52	white
TCGA-92-7340	LUAD	female	45	white
TCGA-BJ-A28R	THCA	female	38	white
TCGA-BJ-A28X	THCA	female	32	white

Table 15 (continued).

TCGA-BJ-A290	THCA	male	70	white
TCGA-BJ-A2N7	THCA	female	30	white
TCGA-BJ-A2N8	THCA	female	30	white
TCGA-BJ-A2N9	THCA	female	42	white
TCGA-BJ-A2NA	THCA	male	77	white
TCGA-BJ-A3PR	THCA	female	70	white
TCGA-BJ-A3PU	THCA	male	52	white
TCGA-DO-A1JZ	THCA	female	23	black or african american
TCGA-E8-A2JQ	THCA	female	18	white
TCGA-EL-A3GZ	THCA	female	34	white
TCGA-EL-A3H1	THCA	female	66	black or african american
TCGA-EL-A3H2	THCA	male	58	white
TCGA-EL-A3H7	THCA	female	36	white
TCGA-EL-A3MW	THCA	female	55	white
TCGA-EL-A3MX	THCA	female	66	white
TCGA-EL-A3MY	THCA	male	81	white
TCGA-EL-A3N3	THCA	female	53	white
TCGA-EL-A3T0	THCA	female	46	white
TCGA-EL-A3T1	THCA	female	38	black or african american
TCGA-EL-A3T2	THCA	female	55	white
TCGA-EL-A3T3	THCA	male	63	white
TCGA-EL-A3T7	THCA	female	47	white
TCGA-EL-A3T8	THCA	male	36	asian
TCGA-EL-A3TA	THCA	male	42	white
TCGA-EL-A3TB	THCA	female	47	white
TCGA-EL-A3ZG	THCA	male	15	asian
TCGA-EL-A3ZH	THCA	female	42	white
TCGA-EL-A3ZK	THCA	female	41	white
TCGA-EL-A3ZL	THCA	female	34	white
TCGA-EL-A3ZM	THCA	male	60	not reported
TCGA-EL-A3ZO	THCA	female	79	white
TCGA-EL-A3ZP	THCA	male	19	white
TCGA-EL-A3ZR	THCA	female	46	white
TCGA-EL-A3ZS	THCA	female	22	white
TCGA-EL-A3ZT	THCA	male	35	white
TCGA-EM-A1CS	THCA	female	55	not reported
TCGA-EM-A1CT	THCA	male	76	not reported
TCGA-EM-A1CU	THCA	male	31	not reported
TCGA-EM-A1CV	THCA	female	32	not reported
TCGA-EM-A1CW	THCA	female	39	not reported
TCGA-EM-A1YC	THCA	female	71	not reported

Table 15 (continued).

TCGA-ET-A2MX	THCA	male	27	white
TCGA-ET-A2N5	THCA	female	46	black or african american
TCGA-ET-A3DP	THCA	female	43	black or african american
TCGA-ET-A3DW	THCA	male	64	asian
TCGA-FY-A3TY	THCA	female	61	white
TCGA-GE-A2C6	THCA	female	33	white
TCGA-H2-A2K9	THCA	male	25	white
TCGA-KS-A41I	THCA	female	47	white
TCGA-KS-A41J	THCA	female	28	white
TCGA-KS-A41L	THCA	female	39	white

APPENDIX D.

SUPPLEMENTARY INFORMATION FOR CHAPTER 5

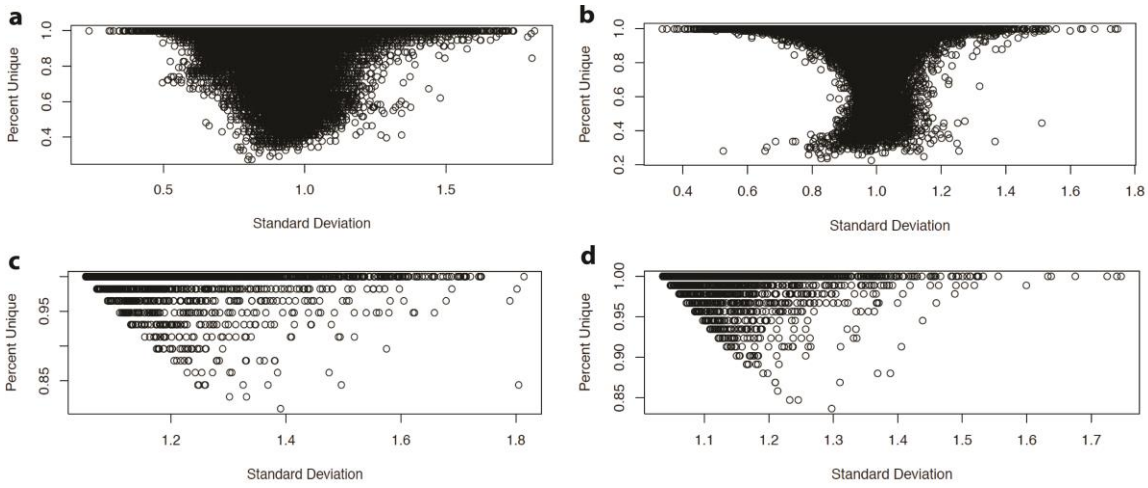

Figure 43 Dimension reduction of genes.

Table 27 Genes selected by random forest variable importance.

	Ensemble gene ID	Pathway ID	Pathway Name
Gemcitabine	ENSG00000137275	P00006	Apoptosis signaling pathway
	ENSG00000169413	P00005	Angiogenesis
	ENSG00000067900	P00016	Cytoskeletal regulation by Rho GTPase
	ENSG00000067900	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000067900	P06959	CCKR signaling map
	ENSG00000187764	P00007	Axon guidance mediated by semaphorins
	ENSG00000101144	P00052	TGF-beta signaling pathway
	ENSG00000101144	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000079215	P00037	Ionotropic glutamate receptor pathway
	ENSG00000079215	P00039	Metabotropic glutamate receptor group III pathway
	ENSG00000130821	P00043	Muscarinic acetylcholine receptor 2 and 4 signaling pathway
	ENSG00000130821	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000080503	P00057	Wnt signaling pathway
	ENSG00000099956	P00057	Wnt signaling pathway
	ENSG00000275837	P00057	Wnt signaling pathway
	ENSG00000073584	P00057	Wnt signaling pathway
	ENSG00000132639	P00001	Adrenaline and noradrenaline biosynthesis
	ENSG00000132639	P00002	Alpha adrenergic receptor signaling pathway
	ENSG00000132639	P00037	Ionotropic glutamate receptor pathway
	ENSG00000132639	P00039	Metabotropic glutamate receptor group III pathway
	ENSG00000132639	P00040	Metabotropic glutamate receptor group II pathway
	ENSG00000132639	P00042	Muscarinic acetylcholine receptor 1 and 3 signaling pathway
	ENSG00000132639	P00043	Muscarinic acetylcholine receptor 2 and 4 signaling pathway
	ENSG00000132639	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000132639	P04373	5HT1 type receptor mediated signaling pathway
	ENSG00000132639	P04374	5HT2 type receptor mediated signaling pathway
	ENSG00000132639	P04375	5HT3 type receptor mediated signaling pathway
	ENSG00000132639	P04376	5HT4 type receptor mediated signaling pathway
	ENSG00000132639	P04377	Beta1 adrenergic receptor signaling pathway
	ENSG00000132639	P04378	Beta2 adrenergic receptor signaling pathway
	ENSG00000132639	P04379	Beta3 adrenergic receptor signaling pathway
	ENSG00000132639	P04380	Cortocotropin releasing factor receptor signaling pathway
	ENSG00000132639	P04391	Oxytocin receptor mediated signaling pathway
	ENSG00000132639	P04394	Thyrotropin-releasing hormone receptor signaling pathway

Table 28 (continued).

	ENSG00000132639	P05734	Synaptic_vesicle_trafficking
	ENSG00000132639	P05912	Dopamine receptor mediated signaling pathway
	ENSG00000132639	P05915	Opioid proenkephalin pathway
	ENSG00000132639	P05916	Opioid prodynorphin pathway
	ENSG00000132639	P05917	Opioid proopiomelanocortin pathway
	ENSG00000057252	P02727	Androgen/estrogene/progesterone biosynthesis
	ENSG00000167780	P02727	Androgen/estrogene/progesterone biosynthesis
	ENSG00000100485	P00005	Angiogenesis
	ENSG00000100485	P00010	B cell activation
	ENSG00000100485	P00018	EGF receptor signaling pathway
	ENSG00000100485	P00021	FGF signaling pathway
	ENSG00000100485	P00032	Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade
	ENSG00000100485	P00034	Integrin signalling pathway
	ENSG00000100485	P00036	Interleukin signaling pathway
	ENSG00000100485	P00047	PDGF signaling pathway
	ENSG00000100485	P00048	PI3 kinase pathway
	ENSG00000100485	P00053	T cell activation
	ENSG00000100485	P04393	Ras Pathway
	ENSG00000100485	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000066336	P00036	Interleukin signaling pathway
	ENSG00000104549	P00014	Cholesterol biosynthesis
	ENSG00000010671	P00010	B cell activation
	ENSG00000278195	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000278195	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000138378	P00018	EGF receptor signaling pathway
	ENSG00000138378	P00036	Interleukin signaling pathway
	ENSG00000138378	P00038	JAK/STAT signaling pathway
	ENSG00000138378	P00047	PDGF signaling pathway
	ENSG00000126561	P00018	EGF receptor signaling pathway
	ENSG00000126561	P00036	Interleukin signaling pathway
	ENSG00000126561	P00038	JAK/STAT signaling pathway
	ENSG00000126561	P00047	PDGF signaling pathway
	ENSG00000079950	P00001	Adrenaline and noradrenaline biosynthesis
	ENSG00000079950	P00049	Parkinson disease
	ENSG00000196628	P06959	CCKR signaling map
	ENSG00000148737	P00004	Alzheimer disease-presenilin pathway
	ENSG00000148737	P00005	Angiogenesis

Table 29 (continued).

	ENSG00000148737	P00012	Cadherin signaling pathway
	ENSG00000148737	P00057	Wnt signaling pathway
	ENSG00000148516	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000028137	P00006	Apoptosis signaling pathway
	ENSG00000133107	P00004	Alzheimer disease-presenilin pathway
	ENSG00000105397	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000170142	P00060	Ubiquitin proteasome pathway
	ENSG00000154277	P00049	Parkinson disease
	ENSG00000141968	P00010	B cell activation
	ENSG00000141968	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000141968	P00047	PDGF signaling pathway
	ENSG00000141968	P00053	T cell activation
	ENSG00000015285	P00053	T cell activation
	ENSG00000106299	P00016	Cytoskeletal regulation by Rho GTPase
	ENSG00000106299	P00029	Huntington disease
	ENSG00000115085	P00053	T cell activation
	ENSG00000171700	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000171700	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000136286	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000067191	P00003	Alzheimer disease-amyloid secretase pathway
	ENSG00000067191	P00039	Metabotropic glutamate receptor group III pathway
	ENSG00000067191	P00040	Metabotropic glutamate receptor group II pathway
	ENSG00000067191	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000067191	P04374	5HT2 type receptor mediated signaling pathway
	ENSG00000067191	P04377	Beta1 adrenergic receptor signaling pathway
	ENSG00000067191	P04378	Beta2 adrenergic receptor signaling pathway
	ENSG00000067191	P04391	Oxytocin receptor mediated signaling pathway
	ENSG00000067191	P04394	Thyrotropin-releasing hormone receptor signaling pathway
	ENSG00000110031	P00005	Angiogenesis
	ENSG00000110031	P00056	VEGF signaling pathway
	ENSG00000123329	P00047	PDGF signaling pathway
	ENSG00000008516	P00004	Alzheimer disease-presenilin pathway
	ENSG00000152689	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000149260	P00029	Huntington disease
	ENSG00000170075	P00049	Parkinson disease

Table 30 (continued).

	ENSG00000118514	P04372	5-Hydroxytryptamine degredation
	ENSG00000164674	P05734	Synaptic_vesicle_trafficking
	ENSG00000196664	P00054	Toll receptor signaling pathway
	ENSG00000174123	P00054	Toll receptor signaling pathway
	ENSG00000166664	P00003	Alzheimer disease-amyloid secretase pathway
	ENSG00000166664	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000275917	P00003	Alzheimer disease-amyloid secretase pathway
	ENSG00000275917	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000118971	P00013	Cell cycle
	ENSG00000118971	P00048	PI3 kinase pathway
	ENSG00000112576	P00013	Cell cycle
	ENSG00000163823	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000121807	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000183813	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000112486	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000173585	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000100024	P02771	Pyrimidine Metabolism
	ENSG00000012124	P00010	B cell activation
	ENSG00000178562	P00053	T cell activation
	ENSG00000172215	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000004468	P06959	CCKR signaling map
	ENSG00000167286	P00053	T cell activation
	ENSG00000198851	P00053	T cell activation
	ENSG00000160654	P00053	T cell activation
	ENSG00000198821	P00053	T cell activation
	ENSG00000198373	P00060	Ubiquitin proteasome pathway
	ENSG00000019582	P00053	T cell activation
	ENSG00000166225	P00005	Angiogenesis
	ENSG00000166225	P00021	FGF signaling pathway
	ENSG00000105369	P00010	B cell activation
	ENSG00000007312	P00010	B cell activation
	ENSG00000062598	P00034	Integrin signalling pathway
	ENSG00000160219	P00018	EGF receptor signaling pathway
	ENSG00000140044	P00006	Apoptosis signaling pathway
	ENSG00000113361	P00012	Cadherin signaling pathway

Table 31 (continued).

	ENSG00000113361	P00057	Wnt signaling pathway
	ENSG00000144837	P05726	2-arachidonoylglycerol biosynthesis
	ENSG00000100568	P00001	Adrenaline and noradrenaline biosynthesis
	ENSG00000198001	P00054	Toll receptor signaling pathway
	ENSG00000099365	P00039	Metabotropic glutamate receptor group III pathway
	ENSG00000099365	P00040	Metabotropic glutamate receptor group II pathway
	ENSG00000099365	P00042	Muscarinic acetylcholine receptor 1 and 3 signaling pathway
	ENSG00000099365	P00043	Muscarinic acetylcholine receptor 2 and 4 signaling pathway
	ENSG00000099365	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000099365	P05734	Synaptic_vesicle_trafficking
	ENSG00000213658	P00053	T cell activation
	ENSG00000104951	P04372	5-Hydroxytryptamine degredation
	ENSG00000132718	P05734	Synaptic_vesicle_trafficking
	ENSG00000114737	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000114737	P00035	Interferon-gamma signaling pathway
	ENSG00000196405	P00016	Cytoskeletal regulation by Rho GTPase
	ENSG00000189007	P02739	De novo pyrimidine deoxyribonucleotide biosynthesis
	ENSG00000189007	P02774	Salvage pyrimidine deoxyribonucleotides
	ENSG00000189007	P02775	Salvage pyrimidine ribonucleotides
	ENSG00000137171	P00003	Alzheimer disease-amyloid secretase pathway
	ENSG00000136280	P00018	EGF receptor signaling pathway
	ENSG00000197410	P00012	Cadherin signaling pathway
	ENSG00000197410	P00057	Wnt signaling pathway
	ENSG00000284227	P00012	Cadherin signaling pathway
	ENSG00000284227	P00057	Wnt signaling pathway
	ENSG00000165323	P00012	Cadherin signaling pathway
	ENSG00000165323	P00057	Wnt signaling pathway
	ENSG00000282908	P00012	Cadherin signaling pathway
	ENSG00000282908	P00057	Wnt signaling pathway
	ENSG00000169118	P00049	Parkinson disease
	ENSG00000169118	P00057	Wnt signaling pathway
	ENSG00000146094	P00005	Angiogenesis
	ENSG00000278259	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000278372	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000128271	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha

Table 32 (continued).

	ENSG00000128271	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000123454	P00001	Adrenaline and noradrenaline biosynthesis
	ENSG00000123454	P05912	Dopamine receptor mediated signaling pathway
	ENSG00000276231	P00018	EGF receptor signaling pathway
	ENSG00000276231	P00019	Endothelin signaling pathway
	ENSG00000120907	P00002	Alpha adrenergic receptor signaling pathway
	ENSG00000120907	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000150594	P00002	Alpha adrenergic receptor signaling pathway
	ENSG00000150594	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000180096	P00049	Parkinson disease
	ENSG00000118363	P04395	Vasopressin synthesis
	ENSG00000239900	P02738	De novo purine biosynthesis
	ENSG00000114841	P00029	Huntington disease
	ENSG00000140795	P00016	Cytoskeletal regulation by Rho GTPase
	ENSG00000140795	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000172667	P00059	p53 pathway
	ENSG00000184845	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000184845	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000184845	P05912	Dopamine receptor mediated signaling pathway
	ENSG00000158050	P00046	Oxidative stress response
	ENSG00000128951	P02739	De novo pyrimidine deoxyribonucleotide biosynthesis
	ENSG00000161202	P00004	Alzheimer disease-presenilin pathway
	ENSG00000161202	P00005	Angiogenesis
	ENSG00000161202	P00057	Wnt signaling pathway
	ENSG00000124126	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000111674	P00024	Glycolysis
	ENSG00000088367	P05912	Dopamine receptor mediated signaling pathway
	ENSG00000088367	P06587	Nicotine pharmacodynamics pathway
	ENSG00000181104	P00005	Angiogenesis
	ENSG00000181104	P00011	Blood coagulation
	ENSG00000137714	P04396	Vitamin D metabolism and pathway
	ENSG00000129682	P00021	FGF signaling pathway
	ENSG00000000938	P00049	Parkinson disease

ENSG00000187474	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
ENSG00000010810	P00007	Axon guidance mediated by semaphorins
ENSG00000010810	P00012	Cadherin signaling pathway
ENSG00000010810	P00034	Integrin signalling pathway
ENSG00000010810	P00049	Parkinson disease
ENSG00000204681	P05731	GABA-B_receptor_II_signaling
ENSG00000206466	P05731	GABA-B_receptor_II_signaling
ENSG00000206511	P05731	GABA-B_receptor_II_signaling
ENSG00000232569	P05731	GABA-B_receptor_II_signaling
ENSG00000232632	P05731	GABA-B_receptor_II_signaling
ENSG00000237051	P05731	GABA-B_receptor_II_signaling
ENSG00000237112	P05731	GABA-B_receptor_II_signaling
ENSG00000154727	P00047	PDGF signaling pathway
ENSG00000258643	P00006	Apoptosis signaling pathway
ENSG00000132965	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
ENSG00000112699	P02752	Mannose metabolism
ENSG00000167083	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
ENSG00000167083	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
ENSG00000167083	P00028	Heterotrimeric G-protein signaling pathway-rod outer segment phototransduction
ENSG00000167083	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
ENSG00000167083	P00039	Metabotropic glutamate receptor group III pathway
ENSG00000167083	P00040	Metabotropic glutamate receptor group II pathway
ENSG00000167083	P00042	Muscarinic acetylcholine receptor 1 and 3 signaling pathway
ENSG00000167083	P00043	Muscarinic acetylcholine receptor 2 and 4 signaling pathway
ENSG00000167083	P00048	PI3 kinase pathway
ENSG00000167083	P00057	Wnt signaling pathway
ENSG00000167083	P04373	5HT1 type receptor mediated signaling pathway
ENSG00000167083	P04374	5HT2 type receptor mediated signaling pathway
ENSG00000167083	P04376	5HT4 type receptor mediated signaling pathway
ENSG00000167083	P04377	Beta1 adrenergic receptor signaling pathway
ENSG00000167083	P04378	Beta2 adrenergic receptor signaling pathway
ENSG00000167083	P04379	Beta3 adrenergic receptor signaling pathway
ENSG00000167083	P04380	Cortocotropin releasing factor receptor signaling pathway
ENSG00000167083	P04385	Histamine H1 receptor mediated signaling pathway
ENSG00000167083	P04386	Histamine H2 receptor mediated signaling pathway

Table 33 (continued).

	ENSG00000167083	P04391	Oxytocin receptor mediated signaling pathway
	ENSG00000167083	P04394	Thyrotropin-releasing hormone receptor signaling pathway
	ENSG00000167083	P05730	Endogenous_cannabinoid_signaling
	ENSG00000167083	P05731	GABA-B_receptor_II_signaling
	ENSG00000167083	P05911	Angiotensin II-stimulated signaling through G proteins and beta-arrestin
	ENSG00000167083	P05913	Enkephalin release
	ENSG00000167083	P05915	Opioid proenkephalin pathway
	ENSG00000167083	P05916	Opioid prodynorphin pathway
	ENSG00000167083	P05917	Opioid proopiomelanocortin pathway
	ENSG00000186810	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000198055	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000198055	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000198055	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000198055	P05911	Angiotensin II-stimulated signaling through G proteins and beta-arrestin
	ENSG00000177885	P00005	Angiogenesis
	ENSG00000177885	P00010	B cell activation
	ENSG00000177885	P00018	EGF receptor signaling pathway
	ENSG00000177885	P00021	FGF signaling pathway
	ENSG00000177885	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000177885	P00032	Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade
	ENSG00000177885	P00034	Integrin signalling pathway
	ENSG00000177885	P00036	Interleukin signaling pathway
	ENSG00000177885	P00047	PDGF signaling pathway
	ENSG00000177885	P00048	PI3 kinase pathway
	ENSG00000177885	P00053	T cell activation
	ENSG00000177885	P04393	Ras Pathway
	ENSG00000177885	P05912	Dopamine receptor mediated signaling pathway
	ENSG00000177885	P06587	Nicotine pharmacodynamics pathway
	ENSG00000177885	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000177885	P06959	CCKR signaling map
	ENSG00000164418	P00029	Huntington disease
	ENSG00000164418	P00037	Ionotropic glutamate receptor pathway

Table 34 (continued).

	ENSG00000164418	P00039	Metabotropic glutamate receptor group III pathway
	ENSG00000179603	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000179603	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000179603	P00039	Metabotropic glutamate receptor group III pathway
	ENSG00000163739	P06959	CCKR signaling map
	ENSG00000152402	P00019	Endothelin signaling pathway
	ENSG00000143774	P02738	De novo purine biosynthesis
	ENSG00000101336	P00049	Parkinson disease
	ENSG00000168384	P00053	T cell activation
	ENSG00000206291	P00053	T cell activation
	ENSG00000224103	P00053	T cell activation
	ENSG00000228163	P00053	T cell activation
	ENSG00000229685	P00053	T cell activation
	ENSG00000231389	P00053	T cell activation
	ENSG00000235844	P00053	T cell activation
	ENSG00000236177	P00053	T cell activation
	ENSG00000196735	P00053	T cell activation
	ENSG00000206305	P00053	T cell activation
	ENSG00000225890	P00053	T cell activation
	ENSG00000228284	P00053	T cell activation
	ENSG00000232062	P00053	T cell activation
	ENSG00000236418	P00053	T cell activation
	ENSG00000204287	P00053	T cell activation
	ENSG00000206308	P00053	T cell activation
	ENSG00000226260	P00053	T cell activation
	ENSG00000227993	P00053	T cell activation
	ENSG00000228987	P00053	T cell activation
	ENSG00000230726	P00053	T cell activation
	ENSG00000234794	P00053	T cell activation
	ENSG00000277263	P00053	T cell activation
	ENSG00000189403	P00059	p53 pathway
	ENSG00000086696	P02727	Androgen/estrogene/progesterone biosynthesis
	ENSG00000211899	P00010	B cell activation
	ENSG00000282657	P00010	B cell activation
	ENSG00000110324	P00036	Interleukin signaling pathway
	ENSG00000095752	P00036	Interleukin signaling pathway
	ENSG00000081985	P00036	Interleukin signaling pathway
	ENSG00000100385	P00036	Interleukin signaling pathway

Table 35 (continued).

	ENSG00000163464	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000163464	P00036	Interleukin signaling pathway
	ENSG00000180871	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000180871	P00036	Interleukin signaling pathway
	ENSG00000171105	P00032	Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade
	ENSG00000171105	P00033	Insulin/IGF pathway-protein kinase B signaling cascade
	ENSG00000171105	P00048	PI3 kinase pathway
	ENSG00000171105	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000156886	P00034	Integrin signalling pathway
	ENSG00000083457	P00034	Integrin signalling pathway
	ENSG00000169896	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000169896	P00034	Integrin signalling pathway
	ENSG00000140678	P00034	Integrin signalling pathway
	ENSG00000160255	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000160255	P00034	Integrin signalling pathway
	ENSG00000123104	P00010	B cell activation
	ENSG00000123104	P00019	Endothelin signaling pathway
	ENSG00000123104	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000123104	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000123104	P00042	Muscarinic acetylcholine receptor 1 and 3 signaling pathway
	ENSG00000123104	P00047	PDGF signaling pathway
	ENSG00000123104	P00057	Wnt signaling pathway
	ENSG00000123104	P04385	Histamine H1 receptor mediated signaling pathway
	ENSG00000123104	P05911	Angiotensin II-stimulated signaling through G proteins and beta-arrestin
	ENSG00000123104	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000101384	P00005	Angiogenesis
	ENSG00000101384	P00045	Notch signaling pathway
	ENSG00000120457	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000120457	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha

Table 36 (continued).

	ENSG00000120457	P00043	Muscarinic acetylcholine receptor 2 and 4 signaling pathway
	ENSG00000143761	P00029	Huntington disease
	ENSG00000143761	P00034	Integrin signalling pathway
	ENSG00000043462	P00053	T cell activation
	ENSG00000165527	P00029	Huntington disease
	ENSG00000165527	P00034	Integrin signalling pathway
	ENSG00000107798	P02727	Androgen/estrogene/progesterone biosynthesis
	ENSG00000079435	P02782	Triacylglycerol metabolism
	ENSG00000155366	P00005	Angiogenesis
	ENSG00000155366	P00007	Axon guidance mediated by semaphorins
	ENSG00000155366	P00008	Axon guidance mediated by Slit/Robo
	ENSG00000155366	P00016	Cytoskeletal regulation by Rho GTPase
	ENSG00000155366	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000155366	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000155366	P00034	Integrin signalling pathway
	ENSG00000155366	P04393	Ras Pathway
	ENSG00000155366	P05911	Angiotensin II-stimulated signaling through G proteins and beta-arrestin
	ENSG00000070018	P00004	Alzheimer disease-presenilin pathway
	ENSG00000070018	P00057	Wnt signaling pathway
	ENSG00000281324	P00004	Alzheimer disease-presenilin pathway
	ENSG00000281324	P00057	Wnt signaling pathway
	ENSG00000204487	P00006	Apoptosis signaling pathway
	ENSG00000206437	P00006	Apoptosis signaling pathway
	ENSG00000223448	P00006	Apoptosis signaling pathway
	ENSG00000227507	P00006	Apoptosis signaling pathway
	ENSG00000231314	P00006	Apoptosis signaling pathway
	ENSG00000236237	P00006	Apoptosis signaling pathway
	ENSG00000236925	P00006	Apoptosis signaling pathway
	ENSG00000238114	P00006	Apoptosis signaling pathway
	ENSG00000137834	P00052	TGF-beta signaling pathway
	ENSG00000006062	P00006	Apoptosis signaling pathway
	ENSG00000006062	P00018	EGF receptor signaling pathway
	ENSG00000006062	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000006062	P06959	CCKR signaling map
	ENSG00000282637	P00006	Apoptosis signaling pathway
	ENSG00000282637	P00018	EGF receptor signaling pathway

Table 37 (continued).

	ENSG00000282637	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000282637	P06959	CCKR signaling map
	ENSG00000198909	P00010	B cell activation
	ENSG00000198909	P00018	EGF receptor signaling pathway
	ENSG00000198909	P00021	FGF signaling pathway
	ENSG00000198909	P00034	Integrin signalling pathway
	ENSG00000198909	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000104814	P00006	Apoptosis signaling pathway
	ENSG00000104814	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000282928	P00006	Apoptosis signaling pathway
	ENSG00000282928	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000198625	P00033	Insulin/IGF pathway-protein kinase B signaling cascade
	ENSG00000198625	P00059	p53 pathway
	ENSG00000198625	P04392	P53 pathway feedback loops 1
	ENSG00000198625	P04398	p53 pathway feedback loops 2
	ENSG00000137486	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000137486	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000137486	P00057	Wnt signaling pathway
	ENSG00000137486	P05911	Angiotensin II-stimulated signaling through G proteins and beta-arrestin
	ENSG00000141480	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000141480	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000141480	P00057	Wnt signaling pathway
	ENSG00000141480	P05911	Angiotensin II-stimulated signaling through G proteins and beta-arrestin
	ENSG00000141480	P06959	CCKR signaling map
	ENSG00000215914	P00004	Alzheimer disease-presenilin pathway
	ENSG00000167508	P00014	Cholesterol biosynthesis
	ENSG00000070669	P02730	Asparagine and aspartate biosynthesis
	ENSG00000091536	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000128641	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000176658	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000142347	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000197535	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000169994	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000099331	P00044	Nicotinic acetylcholine receptor signaling pathway

Table 38 (continued).

	ENSG00000107954	P00045	Notch signaling pathway
	ENSG00000196712	P00018	EGF receptor signaling pathway
	ENSG00000101096	P00009	Axon guidance mediated by netrin
	ENSG00000101096	P00010	B cell activation
	ENSG00000101096	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000101096	P00053	T cell activation
	ENSG00000101096	P00057	Wnt signaling pathway
	ENSG00000101096	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000101096	P06959	CCKR signaling map
	ENSG00000007171	P00048	PI3 kinase pathway
	ENSG00000099250	P00007	Axon guidance mediated by semaphorins
	ENSG00000133961	P00045	Notch signaling pathway
	ENSG00000125510	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000125510	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000277044	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000277044	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000124507	P00029	Huntington disease
	ENSG00000113555	P00012	Cadherin signaling pathway
	ENSG00000113555	P00057	Wnt signaling pathway
	ENSG00000255408	P00012	Cadherin signaling pathway
	ENSG00000255408	P00057	Wnt signaling pathway
	ENSG00000204967	P00012	Cadherin signaling pathway
	ENSG00000204967	P00057	Wnt signaling pathway
	ENSG00000204965	P00012	Cadherin signaling pathway
	ENSG00000204965	P00057	Wnt signaling pathway
	ENSG00000204963	P00012	Cadherin signaling pathway
	ENSG00000204963	P00057	Wnt signaling pathway
	ENSG00000187372	P00012	Cadherin signaling pathway
	ENSG00000187372	P00057	Wnt signaling pathway
	ENSG00000120327	P00012	Cadherin signaling pathway
	ENSG00000120327	P00057	Wnt signaling pathway
	ENSG00000113212	P00012	Cadherin signaling pathway
	ENSG00000113212	P00057	Wnt signaling pathway
	ENSG00000253846	P00012	Cadherin signaling pathway
	ENSG00000253846	P00057	Wnt signaling pathway

Table 39 (continued).

	ENSG00000253767	P00012	Cadherin signaling pathway
	ENSG00000253767	P00057	Wnt signaling pathway
	ENSG00000185527	P00028	Heterotrimeric G-protein signaling pathway-rod outer segment phototransduction
	ENSG00000141959	P00024	Glycolysis
	ENSG00000171608	P00005	Angiogenesis
	ENSG00000171608	P00006	Apoptosis signaling pathway
	ENSG00000171608	P00009	Axon guidance mediated by netrin
	ENSG00000171608	P00010	B cell activation
	ENSG00000171608	P00018	EGF receptor signaling pathway
	ENSG00000171608	P00019	Endothelin signaling pathway
	ENSG00000171608	P00021	FGF signaling pathway
	ENSG00000171608	P00030	Hypoxia response via HIF activation
	ENSG00000171608	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000171608	P00033	Insulin/IGF pathway-protein kinase B signaling cascade
	ENSG00000171608	P00034	Integrin signalling pathway
	ENSG00000171608	P00047	PDGF signaling pathway
	ENSG00000171608	P00053	T cell activation
	ENSG00000171608	P00056	VEGF signaling pathway
	ENSG00000171608	P00059	p53 pathway
	ENSG00000171608	P04393	Ras Pathway
	ENSG00000171608	P04398	p53 pathway feedback loops 2
	ENSG00000105499	P00005	Angiogenesis
	ENSG00000105499	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000101333	P00002	Alpha adrenergic receptor signaling pathway
	ENSG00000101333	P00019	Endothelin signaling pathway
	ENSG00000101333	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000101333	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000101333	P00041	Metabotropic glutamate receptor group I pathway
	ENSG00000101333	P00042	Muscarinic acetylcholine receptor 1 and 3 signaling pathway
	ENSG00000101333	P00057	Wnt signaling pathway
	ENSG00000101333	P04374	5HT2 type receptor mediated signaling pathway
	ENSG00000101333	P04385	Histamine H1 receptor mediated signaling pathway
	ENSG00000101333	P04391	Oxytocin receptor mediated signaling pathway
	ENSG00000101333	P04394	Thyrotropin-releasing hormone receptor signaling pathway

Table 40 (continued).

	ENSG00000108387	P00049	Parkinson disease
	ENSG00000120910	P00010	B cell activation
	ENSG00000120910	P00053	T cell activation
	ENSG00000120910	P00057	Wnt signaling pathway
	ENSG00000163932	P00002	Alpha adrenergic receptor signaling pathway
	ENSG00000163932	P00003	Alzheimer disease-amyloid secretase pathway
	ENSG00000163932	P00005	Angiogenesis
	ENSG00000163932	P00006	Apoptosis signaling pathway
	ENSG00000163932	P00010	B cell activation
	ENSG00000163932	P00018	EGF receptor signaling pathway
	ENSG00000163932	P00019	Endothelin signaling pathway
	ENSG00000163932	P00021	FGF signaling pathway
	ENSG00000163932	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000163932	P00042	Muscarinic acetylcholine receptor 1 and 3 signaling pathway
	ENSG00000163932	P00056	VEGF signaling pathway
	ENSG00000163932	P00057	Wnt signaling pathway
	ENSG00000163932	P04374	5HT2 type receptor mediated signaling pathway
	ENSG00000163932	P04385	Histamine H1 receptor mediated signaling pathway
	ENSG00000163932	P04391	Oxytocin receptor mediated signaling pathway
	ENSG00000163932	P04394	Thyrotropin-releasing hormone receptor signaling pathway
	ENSG00000163932	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000163932	P06959	CCKR signaling map
	ENSG00000181790	P00059	p53 pathway
	ENSG00000101182	P00049	Parkinson disease
	ENSG00000120899	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000120899	P00034	Integrin signalling pathway
	ENSG00000120899	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000120899	P06959	CCKR signaling map
	ENSG00000111679	P00005	Angiogenesis
	ENSG00000111679	P00010	B cell activation
	ENSG00000111679	P00021	FGF signaling pathway
	ENSG00000111679	P00035	Interferon-gamma signaling pathway
	ENSG00000081237	P00010	B cell activation
	ENSG00000081237	P00038	JAK/STAT signaling pathway
	ENSG00000081237	P00053	T cell activation
	ENSG00000262418	P00010	B cell activation

Table 41 (continued).

	ENSG00000262418	P00038	JAK/STAT signaling pathway
	ENSG00000262418	P00053	T cell activation
	ENSG00000128340	P00007	Axon guidance mediated by semaphorins
	ENSG00000128340	P00008	Axon guidance mediated by Slit/Robo
	ENSG00000128340	P00009	Axon guidance mediated by netrin
	ENSG00000128340	P00010	B cell activation
	ENSG00000128340	P00016	Cytoskeletal regulation by Rho GTPase
	ENSG00000128340	P00018	EGF receptor signaling pathway
	ENSG00000128340	P00021	FGF signaling pathway
	ENSG00000128340	P00029	Huntington disease
	ENSG00000128340	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000128340	P00034	Integrin signalling pathway
	ENSG00000128340	P00053	T cell activation
	ENSG00000128340	P00056	VEGF signaling pathway
	ENSG00000128340	P04393	Ras Pathway
	ENSG00000128340	P05918	p38 MAPK pathway
	ENSG00000172575	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000171791	P00006	Apoptosis signaling pathway
	ENSG00000171791	P00046	Oxidative stress response
	ENSG00000171791	P06959	CCKR signaling map
	ENSG00000090104	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000090104	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000090104	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
	ENSG00000127074	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha
	ENSG00000127074	P00027	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha
	ENSG00000127074	P00031	Inflammation mediated by chemokine and cytokine signaling pathway
5-Flourouracil	ENSG00000187122	P00008	Axon guidance mediated by Slit/Robo
	ENSG00000153147	P00057	Wnt signaling pathway
	ENSG00000139613	P00057	Wnt signaling pathway
	ENSG00000115904	P00005	Angiogenesis
	ENSG00000115904	P00010	B cell activation
	ENSG00000115904	P00018	EGF receptor signaling pathway
	ENSG00000115904	P00021	FGF signaling pathway

Table 42 (continued).

	ENSG00000115904	P00031	Inflammation mediated by chemokine and
	ENSG00000115904	P00032	Insulin/IGF pathway-mitogen activated protein kinase
	ENSG00000115904	P00034	Integrin signalling pathway
	ENSG00000115904	P00036	Interleukin signaling pathway
	ENSG00000115904	P00047	PDGF signaling pathway
	ENSG00000115904	P00048	PI3 kinase pathway
	ENSG00000115904	P00053	T cell activation
	ENSG00000115904	P04393	Ras Pathway
	ENSG00000115904	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000115904	P06959	CCKR signaling map
	ENSG00000070808	P00031	Inflammation mediated by chemokine and
	ENSG00000070808	P00037	Ionotropic glutamate receptor pathway
	ENSG00000110395	P00018	EGF receptor signaling pathway
	ENSG00000101199	P00034	Integrin signalling pathway
	ENSG00000078814	P00016	Cytoskeletal regulation by Rho GTPase
	ENSG00000078814	P00031	Inflammation mediated by chemokine and
	ENSG00000078814	P00044	Nicotinic acetylcholine receptor signaling pathway
	ENSG00000078814	P00057	Wnt signaling pathway
	ENSG00000198276	P02775	Salvage pyrimidine ribonucleotides
	ENSG00000009335	P00060	Ubiquitin proteasome pathway
	ENSG00000145819	P00034	Integrin signalling pathway
	ENSG00000145819	P00047	PDGF signaling pathway
	ENSG00000156650	P00059	p53 pathway
	ENSG00000281813	P00059	p53 pathway
	ENSG00000049618	P00057	Wnt signaling pathway
	ENSG00000172602	P00034	Integrin signalling pathway
	ENSG00000168615	P00003	Alzheimer disease-amyloid secretase pathway
	ENSG00000282230	P00003	Alzheimer disease-amyloid secretase pathway
	ENSG00000033800	P00035	Interferon-gamma signaling pathway
	ENSG00000033800	P00038	JAK/STAT signaling pathway
	ENSG00000004975	P00004	Alzheimer disease-presenilin pathway
	ENSG00000004975	P00005	Angiogenesis
	ENSG00000004975	P00057	Wnt signaling pathway
	ENSG00000100393	P00026	Heterotrimeric G-protein signaling pathway-Gi alpha
	ENSG00000100393	P00029	Huntington disease
	ENSG00000100393	P00052	TGF-beta signaling pathway
	ENSG00000100393	P00055	Transcription regulation by bZIP transcription
	ENSG00000100393	P00057	Wnt signaling pathway
	ENSG00000100393	P00059	p53 pathway

Table 43 (continued).

	ENSG00000100393	P06211	BMP_signaling_pathway-drosophila
	ENSG00000100393	P06212	DPP-SCW_signaling_pathway
	ENSG00000100393	P06213	DPP_signaling_pathway
	ENSG00000100393	P06214	GBB_signaling_pathway
	ENSG00000100393	P06216	SCW_signaling_pathway
	ENSG00000100393	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000151422	P00012	Cadherin signaling pathway
	ENSG00000213930	P02744	Fructose galactose metabolism
	ENSG00000161905	P00031	Inflammation mediated by chemokine and
	ENSG00000161905	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000105464	P00029	Huntington disease
	ENSG00000105464	P00037	Ionotropic glutamate receptor pathway
	ENSG00000105464	P00039	Metabotropic glutamate receptor group III
	ENSG00000105464	P00041	Metabotropic glutamate receptor group I
	ENSG00000105464	P00042	Muscarinic acetylcholine receptor 1 and
	ENSG00000197386	P00029	Huntington disease
	ENSG00000173110	P00006	Apoptosis signaling pathway
	ENSG00000173110	P00049	Parkinson disease
	ENSG00000120868	P00006	Apoptosis signaling pathway
	ENSG00000120868	P00020	FAS signaling pathway
	ENSG00000120868	P00029	Huntington disease
	ENSG00000120868	P00059	p53 pathway
	ENSG00000185507	P00054	Toll receptor signaling pathway
	ENSG00000276561	P00054	Toll receptor signaling pathway
	ENSG00000169967	P00010	B cell activation
	ENSG00000169967	P00018	EGF receptor signaling pathway
	ENSG00000169967	P00021	FGF signaling pathway
	ENSG00000169967	P00034	Integrin signalling pathway
	ENSG00000169967	P00047	PDGF signaling pathway
	ENSG00000169967	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000120500	P00057	Wnt signaling pathway
	ENSG00000066136	P06664	Gonadotropin releasing hormone receptor pathway
	ENSG00000177463	P06959	CCKR signaling map
	ENSG00000138801	P02778	Sulfate assimilation
	ENSG00000204967	P00012	Cadherin signaling pathway
	ENSG00000204967	P00057	Wnt signaling pathway
	ENSG00000253953	P00012	Cadherin signaling pathway
	ENSG00000253953	P00057	Wnt signaling pathway
	ENSG00000112033	P00057	Wnt signaling pathway
	ENSG00000050820	P00034	Integrin signalling pathway

Table 44 (continued).

	ENSG00000050820	P06959	CCKR signaling map
	ENSG00000285460	P00034	Integrin signalling pathway
	ENSG00000285460	P06959	CCKR signaling map
	ENSG00000035928	P00017	DNA replication

Table 45 Accuracy by cancer type.

5-Fluorouracil Cancer Type	**Accuracy**	**Count**
Colon adenocarcinoma	85.71%	7
Esophageal carcinoma	25.0%	4
Pancreatic adenocarcinoma	90.0%	10
Rectum adenocarcinoma	100.0%	16
Stomach adenocarcinoma	76.0%	25
Gemcitabine Cancer Type	**Accuracy**	**Count**
Bladder Urothelial Carcinoma	100.00%	12
Breast invasive carcinoma	100.0%	3
Cervical squamous cell carcinoma and endocervical adenocarcinoma	100.0%	2
Cholangiocarcinoma	100.0%	7
Head and Neck squamous cell carcinoma	100.0%	2
Liver hepatocellular carcinoma	100.0%	3
Lung adenocarcinoma	100.0%	5
Ovarian serous cystadenocarcinom	75.0%	4
Pancreatic adenocarcinoma	86.0%	43
Pheochromocytoma and Paraganglioma	100.0%	2
Sarcoma	88.9%	9
Skin Cutaneous Melanoma	100.0%	2
Testicular Germ Cell Tumors	100.0%	1
Uterine Corpus Endometrial Carcinoma	100.0%	1

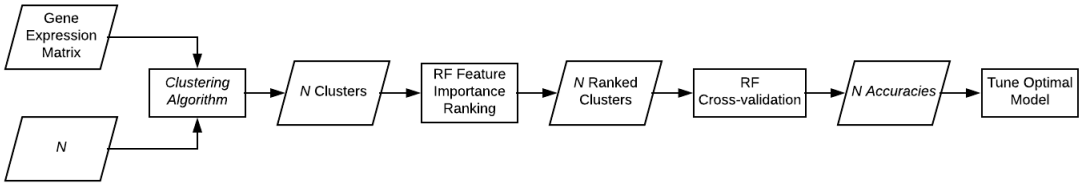

Figure 44 Optimal model workflow.

A pipeline was created for evaluating the performance of models with varying N (number of clusters) for each N between 180 and 220, and clustering algorithms: clara, hierarchical, k-means, model, pam, and sota. All clusters were ranked by a variable importance measure calculated using a random forest classifier. N models were created, where model 1 contained the most informative cluster, and model N contained top N most informative clusters. The mean accuracy for each model was calculated using random forest with 200x cross validation. The model which produced the best accuracy was subsequently tuned for parameter optimization.

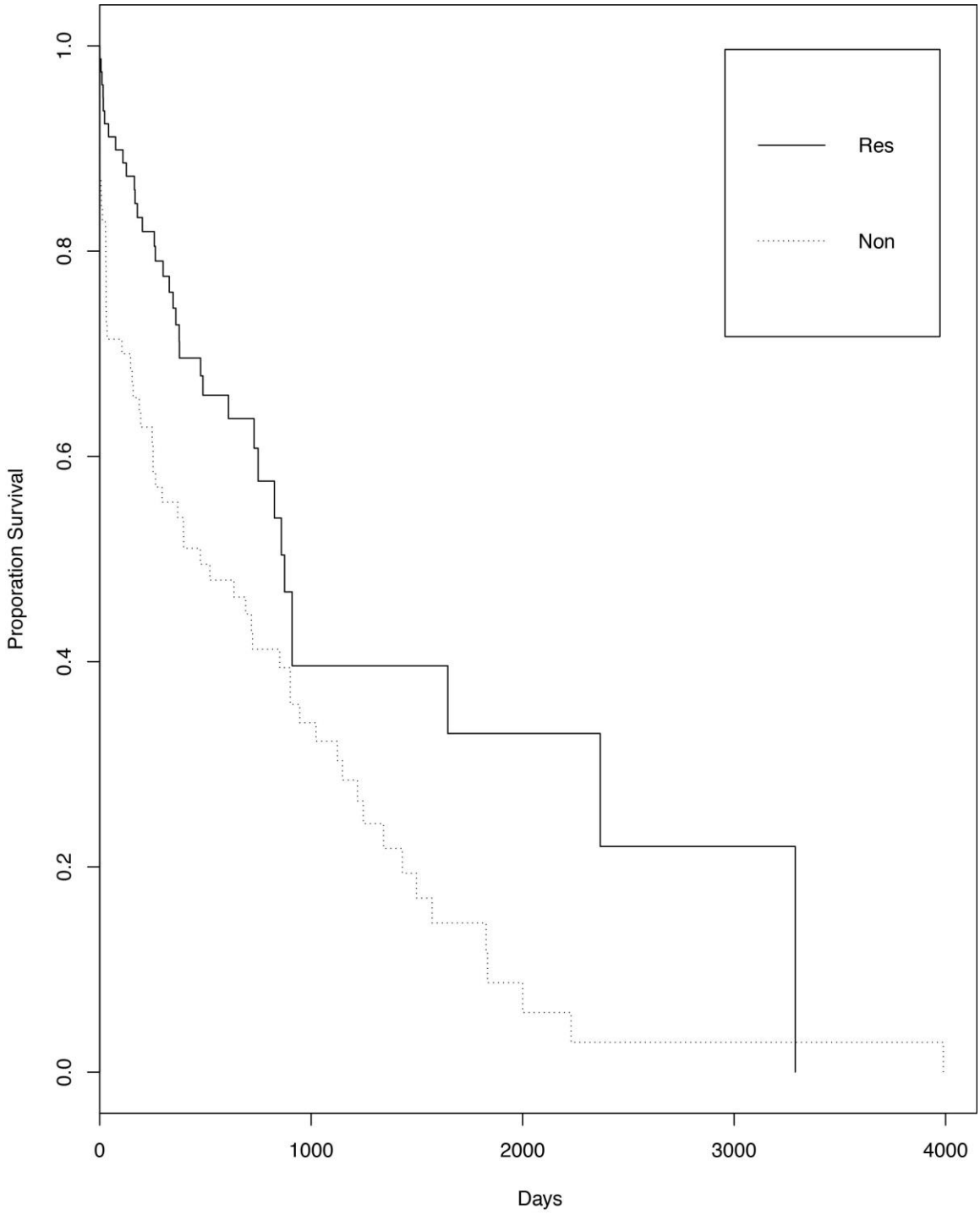

Figure 45 Survival data as a predictor of drug response.

Survival data was downloaded for all patients from both GCB and 5FU models. Kaplan–Meier plot for responders (res) and non-responders (non) depicting patient survival over time.

D.1 Supplementary Methods

D.1.1 Data Pre-Processing

Pre-processing was required to make the data viable for our study. The clinical data was first cleaned, and records were removed if vital data was missing. Records missing drug name, response, cancer type, days on drug therapy, or days to drug therapy start were removed. We also removed records for which gene expression data was not available. We required a sample size of at least 30 records for a drug to be considered for the study and at least 15 records for each response type.

D.1.2 Gene Standardization & Gene Selection

We accounted for the differences of gene expression within cancer types. We calculated the mean (cancer type mean) and standard deviation (cancer type sd) of the logged expression values for every TCGA patient with a given cancer type. We standardized the gene expression for the patients included in the study by subtracting the cancer type mean and dividing by the cancer type standard deviation.

We reduced the number of genes to effectively use clustering algorithms. We wanted to retain the genes that have the highest amount of variation (standard deviation) and the highest number of unique values (Figure 43). Genes with less than 70% unique values across all patients were removed first. Next, we calculated the standard deviation for each remaining gene. We used the product of the standard deviation and percentage of unique values to rank the remaining genes. The top 5,000 genes ranked by this method

were included in the clustering algorithm, which is discussed in the next section. Genes selected by the variable selection method can be found in Table 27.

D.1.3 SPD Results

We compared single cancer and pan-cancer models using Sample Progression Discovery (SPD) clustering and random forest classification. SPD was borrowed from a previously published paper, because it discovers patterns of biological progression within data by building relatively same-size clusters of arbitrary shape. Under SPD, the model cross-validation accuracy was 71.4% for 5-FU and 73.5% for GCB. The number of clusters created by SPD was higher for 5-FU (251) than GCB (202), but the number of clusters selected with random forest was lower (25 vs 35). The results of the single cancer models showed lower accuracy rates than the pan models: 5-FU STAD (70.90%) and GCB PAAD (67.7%). This result supported our decision to only focus on pan-cancer models for the remaining analysis of the study.

D.1.4 Validation Set Investigation

The limited size of the data set was also a bottleneck to performing a robust model validation. We tried two different validation methods. For the first method, we attempted to use the patients who had been treated with 5-FU or GCB but were missing a documented measure of response to treatment. We first proved that the survival data is a good predictor of drug response, and we were going to use survival curves to validate our predictions (See Figure 45 for additional information). Despite its promise, there were too few patients for this approach to be of value. As a second attempt, we used patients who were on combination therapy with the drugs of interest [5-FU/Leucovorin and

Gemcitabine/Cisplatin,] as their first line of treatment. We performed the standard process for model validation (using all training data to train the model, and then testing accuracy on the testing data). The results were not promising [5-FU/Leucovorin: 53.1% and Gemcitabine/Cisplatin: 51.0%]. Thus we left out half of the most populous cancer from the training set as described in the methods.

PUBLICATIONS

1. **Clayton, E.A.**, Khalid, S., Ban, D., Wang, L., Jordan, I.K., and McDonald, J.F. 2019. Tumor suppressor genes and allele-specific expression: Mechanisms and significance. *In Prep.*
2. **Clayton, E.A.**, Rishishwar, L., Huang, T., Gulati, S., Ban, D., McDonald, J.F., and Jordan I.K., 2019. Transposable element induced alternative splicing in cancer. Philosophical Transactions of the Royal Society B. *In Review*.
3. **Clayton, E.A.**, Pujol, T.A., McDonald, J.F., and Qui, P. 2019. Leveraging TCGA gene expression data to build predictive models for cancer drug response. *In Review*.
4. **Clayton, E.A.**, Wang, L., Rishishwar, L., Wang, J., McDonald, J.F., and Jordan, I.K. 2016. Patterns of transposable element expression and insertion in cancer. Frontiers in Molecular Biosciences. 3(76). DOI: 10.3389/fmolb.2016.00076.

5. Huang, C., **Clayton, E.A.**, Matyunina, L.V., McDonald, L.D., Benigno, B.B., Vannberg, F. & McDonald, J.F. 2018. Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. Scientific Reports. 8. DOI: 10.1038/s41598-018-34753-5.
6. Rishishwar, L., Wang, L., **Clayton, E.A.**, Marino-Ramirez, L., McDonald, J.F. and Jordan I.K., 2017. Population and clinical genetics of human transposable elements in the (post) genomic era. Mobile Genetic Elements. 7: 1-20. DOI:10.1080/2159256X.2017.1280116.

REFERENCES

1. McClintock, B., *The origin and behavior of mutable loci in maize.* Proceedings of the National Academy of Sciences, 1950. **36**(6): p. 344-355.

2. Pray, L.A., *Transposons: The jumping genes.* Nature education, 2008. **1**(1): p. 204.

3. Lander, E.S., *Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium.* Nature, 2001. **409**: p. 860-921.

4. Mills, R.E., et al., *Which transposable elements are active in the human genome?* Trends in genetics, 2007. **23**(4): p. 183-191.

5. Hancks, D.C. and H.H. Kazazian Jr, *Active human retrotransposons: variation and disease.* Current opinion in genetics & development, 2012. **22**(3): p. 191-203.

6. Miki, Y., et al., *Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer.* Cancer research, 1992. **52**(3): p. 643-645.

7. Morse, B., et al., *Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma.* Nature, 1988. **333**(6168): p. 87.

8. Lee, E., et al., *Landscape of somatic retrotransposition in human cancers.* Science, 2012. **337**(6097): p. 967-971.

9. Iskow, R.C., et al., *Natural mutagenesis of human genomes by endogenous retrotransposons.* Cell, 2010. **141**(7): p. 1253-1261.

10. Ezkurdia, I., et al., *Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes.* Human Molecular Genetics, 2014. **23**(22): p. 5866-5878.

11. Salzberg, S.L., *Open questions: How many genes do we have?* BMC Biology, 2018. **16**(1): p. 94.

12. Escobar-Hoyos, L., K. Knorr, and O. Abdel-Wahab, *Aberrant RNA Splicing in Cancer.* Annual Review of Cancer Biology, 2019. **3**(1): p. 167-185.

13. Venables, J.P., *Aberrant and alternative splicing in cancer.* Cancer Res, 2004. **64**(21): p. 7647-54.

14. Morrissy, A.S., M. Griffith, and M.A. Marra, *Extensive relationship between antisense transcription and alternative splicing in the human genome.* Genome research, 2011. **21**(8): p. 1203-1212.

15. Du, L. and R.A. Gatti, *Progress toward therapy with antisense-mediated splicing modulation.* Current opinion in molecular therapeutics, 2009. **11**(2): p. 116-123.

16. Mercatante, D.R. and R. Kole, *Control of alternative splicing by antisense oligonucleotides as a potential chemotherapy: effects on gene expression.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2002. **1587**(2): p. 126-132.

17. Sazani, P. and R. Kole, *Therapeutic potential of antisense oligonucleotides as modulators of alternative splicing.* The Journal of Clinical Investigation, 2003. **112**(4): p. 481-486.

18. Guo, X.E., et al., *Targeting tumor suppressor networks for cancer therapeutics.* Curr Drug Targets, 2014. **15**(1): p. 2-16.

19. Chial, H., *Tumor suppressor (TS) genes and the two-hit hypothesis.* Nature Education, 2008. **1**(1): p. 177.

20. Pastinen, T., *Genome-wide allele-specific analysis: insights into regulatory variation.* Nat Rev Genet, 2010. **11**(8): p. 533-8.

21. Shlien, A. and D. Malkin, *Copy number variations and cancer.* Genome medicine, 2009. **1**(6): p. 62-62.

22. Druker, B.J., et al., *Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr–Abl positive cells.* Nature medicine, 1996. **2**(5): p. 561.

23. Tu, S.-M., M.A. Bilen, and N.M. Tannir, *Personalised cancer care: promises and challenges of targeted therapy.* Journal of the Royal Society of Medicine, 2016. **109**(3): p. 98-105.

24. Huang, C., et al., *Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy.* Scientific Reports, 2018. **8**(1): p. 16444.

25. Huang, C., et al., *Open source machine-learning algorithms for the prediction of optimal cancer drug therapies.* PLoS One, 2017. **12**(10): p. e0186906.

26. Vidyasagar, M., *Identifying predictive features in drug response using machine learning: opportunities and challenges.* Annual review of pharmacology and toxicology, 2015. **55**: p. 15-34.

27. de Koning, A.J., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS genetics, 2011. **7**(12): p. e1002384.

28. Wang, H., et al., *SVA elements: a hominid-specific retroposon family.* J Mol Biol, 2005. **354**(4): p. 994-1007.

29. Ostertag, E.M., et al., *SVA elements are nonautonomous retrotransposons that cause disease in humans.* Am J Hum Genet, 2003. **73**(6): p. 1444-51.

30. Kazazian, H.H., Jr., et al., *Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man.* Nature, 1988. **332**(6160): p. 164-6.

31. Brouha, B., et al., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A, 2003. **100**(9): p. 5280-5.

32. Batzer, M.A. and P.L. Deininger, *A human-specific subfamily of Alu sequences.* Genomics, 1991. **9**(3): p. 481-7.

33. Batzer, M.A., et al., *Amplification dynamics of human-specific (HS) Alu family members.* Nucleic Acids Res, 1991. **19**(13): p. 3619-23.

34. Wildschutte, J.H., et al., *Discovery of unfixed endogenous retrovirus insertions in diverse human populations.* Proc Natl Acad Sci U S A, 2016. **113**(16): p. E2326-34.

35. Solyom, S. and H.H. Kazazian, *Mobile elements in the human genome: implications for disease.* Genome medicine, 2012. **4**(2): p. 12.

36. Hancks, D.C. and H.H. Kazazian, Jr., *Active human retrotransposons: variation and disease.* Curr Opin Genet Dev, 2012. **22**(3): p. 191-203.

37. Carreira, P.E., S.R. Richardson, and G.J. Faulkner, *L1 retrotransposons, cancer stem cells and oncogenesis.* The FEBS journal, 2014. **281**(1): p. 63-73.

38. Belancio, V.P., A.M. Roy-Engel, and P.L. Deininger. *All y'all need to know 'bout retroelements in cancer*. in *Seminars in cancer biology*. 2010. Elsevier.

39. Ewing, A.D., *Transposable element detection from whole genome sequence data.* Mob DNA, 2015. **6**: p. 24.

40. Tubio, J.M., et al., *Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes.* Science, 2014. **345**(6196): p. 1251343.

41. Shukla, R., et al., *Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma.* Cell, 2013. **153**(1): p. 101-111.

42. Doucet-O'Hare, T.T., et al., *LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma.* Proceedings of the National Academy of Sciences, 2015. **112**(35): p. E4894-E4900.

43. Scott, E.C., et al., *A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer.* Genome research, 2016. **26**(6): p. 745-755.

44. Kemp, J.R. and M.S. Longworth, *Crossing the LINE Toward Genomic Instability: LINE-1 Retrotransposition in Cancer.* Front Chem, 2015. **3**: p. 68.

45. Belancio, V.P., et al., *Somatic expression of LINE-1 elements in human tissues.* Nucleic Acids Res, 2010. **38**(12): p. 3909-22.

46. Bratthauer, G.L. and T.G. Fanning, *Active LINE-1 retrotransposons in human testicular cancer.* Oncogene, 1992. **7**(3): p. 507-10.

47. Asch, H.L., et al., *Comparative expression of the LINE-1 p40 protein in human breast carcinomas and normal breast tissues.* Oncol Res, 1996. **8**(6): p. 239-47.

48. Bratthauer, G.L. and T.G. Fanning, *LINE-1 retrotransposon expression in pediatric germ cell tumors.* Cancer, 1993. **71**(7): p. 2383-6.

49. Bratthauer, G.L., R.D. Cardiff, and T.G. Fanning, *Expression of LINE-1 retrotransposons in human breast cancer.* Cancer, 1994. **73**(9): p. 2333-6.

50. Doucet-O'Hare, T.T., et al., *Somatically Acquired LINE-1 Insertions in Normal Esophagus Undergo Clonal Expansion in Esophageal Squamous Cell Carcinoma.* Hum Mutat, 2016. **37**(9): p. 942-54.

51. Rodic, N., et al., *Long interspersed element-1 protein expression is a hallmark of many human cancers.* Am J Pathol, 2014. **184**(5): p. 1280-6.

52. Rangasamy, D., et al., *Activation of LINE-1 Retrotransposon Increases the Risk of Epithelial-Mesenchymal Transition and Metastasis in Epithelial Cancer.* Curr Mol Med, 2015. **15**(7): p. 588-97.

53. Stratton, M.R., P.J. Campbell, and P.A. Futreal, *The cancer genome.* Nature, 2009. **458**(7239): p. 719-724.

54. Marx, V., *Cancer genomes: discerning drivers from passengers.* Nature methods, 2014. **11**(4): p. 375-379.

55. Pon, J.R. and M.A. Marra, *Driver and passenger mutations in cancer.* Annu Rev Pathol, 2015. **10**: p. 25-50.

56. Baillie, J.K., et al., *Somatic retrotransposition alters the genetic landscape of the human brain.* Nature, 2011. **479**(7374): p. 534-7.

57. Weinstein, J.N., et al., *The cancer genome atlas pan-cancer analysis project.* Nature genetics, 2013. **45**(10): p. 1113.

58. Jin, Y., et al., *TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets.* Bioinformatics, 2015. **31**(22): p. 3593-9.

59. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.

60. Thung, D.T., et al., *Mobster: accurate detection of mobile element insertions in next generation sequencing data.* Genome Biol, 2014. **15**(10): p. 488.

61. Maltbie, D., L. Ganeshalingam, and P. Allen, *System and method for secure, high-speed transfer of very large files*. 2013, Google Patents.

62. Andrews, S., *FastQC A quality control tool for high throughput sequence data.* Cambridge, UK: Babraham Institute, 2011.

63. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.

64. Pruitt, K.D., et al., *NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy.* Nucleic Acids Res, 2012. **40**(Database issue): p. D130-5.

65. Criscione, S.W., et al., *Transcriptional landscape of repetitive elements in normal and cancer human cells.* BMC Genomics, 2014. **15**: p. 583.

66. Trapnell, C., et al., *Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.* Nat Biotechnol, 2010. **28**(5): p. 511-5.

67. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

68. Penzkofer, T., T. Dandekar, and T. Zemojtel, *L1Base: from functional annotation to prediction of active LINE-1 elements.* Nucleic Acids Res, 2005. **33**(Database issue): p. D498-500.

69. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **43**: p. 11 10 1-33.

70. Rishishwar, L., L. Marino-Ramirez, and I.K. Jordan, *Benchmarking computational tools for polymorphic transposable element detection.* Brief Bioinform, 2016.

71. Quinlan, A.R., *BEDTools: The Swiss-Army Tool for Genome Feature Analysis.* Curr Protoc Bioinformatics, 2014. **47**: p. 11 12 1-34.

72. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2015. **43**(Database issue): p. D805-11.

73. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.

74. Rishishwar, L., C.E. Tellez Villa, and I.K. Jordan, *Transposable element polymorphisms recapitulate human evolution.* Mob DNA, 2015. **6**: p. 21.

75. Deininger, P., *Alu elements: know the SINEs.* Genome Biol, 2011. **12**(12): p. 236.

76. Hancks, D.C. and H.H. Kazazian, Jr., *SVA retrotransposons: Evolution and genetic instability.* Semin Cancer Biol, 2010. **20**(4): p. 234-45.

77. Batzer, M.A. and P.L. Deininger, *Alu repeats and human genomic diversity.* Nat Rev Genet, 2002. **3**(5): p. 370-9.

78. Abbas, S., et al., *Exon 8 splice site mutations in the gene encoding the E3-ligase CBL are associated with core binding factor acute myeloid leukemias.* Haematologica, 2008. **93**(10): p. 1595-7.

79. Aranaz, P., et al., *CBL RING finger deletions are common in core-binding factor acute myeloid leukemias.* Leuk Lymphoma, 2013. **54**(2): p. 428-31.

80. Martinelli, S., et al., *Heterozygous germline mutations in the CBL tumor-suppressor gene cause a Noonan syndrome-like phenotype.* Am J Hum Genet, 2010. **87**(2): p. 250-7.

81. Joshi-Tope, G., et al., *Reactome: a knowledgebase of biological pathways.* Nucleic acids research, 2005. **33**(suppl_1): p. D428-D432.

82. Sever, R. and J.S. Brugge, *Signal transduction in cancer.* Cold Spring Harbor perspectives in medicine, 2015. **5**(4): p. a006098.

83. Damiani, D., et al., *BAALC overexpression retains its negative prognostic role across all cytogenetic risk groups in acute myeloid leukemia patients.* Am J Hematol, 2013. **88**(10): p. 848-52.

84. Zhou, J.D., et al., *Overexpression of BAALC: clinical significance in Chinese de novo acute myeloid leukemia.* Med Oncol, 2015. **32**(1): p. 386.

85. Han, J.S., S.T. Szak, and J.D. Boeke, *Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes.* Nature, 2004. **429**(6989): p. 268-74.

86. Morita, K., et al., *BAALC potentiates oncogenic ERK pathway through interactions with MEKK1 and KLF4.* Leukemia, 2015. **29**(11): p. 2248-56.

87. Harris, C.R., et al., *Association of nuclear localization of a long interspersed nuclear element-1 protein in breast tumors with poor prognostic outcomes.* Genes Cancer, 2010. **1**(2): p. 115-24.

88. de Koning, A.P., et al., *Repetitive elements may comprise over two-thirds of the human genome.* PLoS Genet, 2011. **7**(12): p. e1002384.

89. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

90. Jordan, I.K., et al., *Origin of a substantial fraction of human regulatory sequences from transposable elements.* Trends Genet, 2003. **19**(2): p. 68-72.

91. van de Lagemaat, L.N., et al., *Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions.* Trends Genet, 2003. **19**(10): p. 530-6.

92. Rebollo, R., M.T. Romanish, and D.L. Mager, *Transposable elements: an abundant and natural source of regulatory sequences for host genes.* Annu Rev Genet, 2012. **46**: p. 21-42.

93. Chuong, E.B., N.C. Elde, and C. Feschotte, *Regulatory activities of transposable elements: from conflicts to benefits.* Nat Rev Genet, 2017. **18**(2): p. 71-86.

94. Conley, A.B. and I.K. Jordan, *Identification of transcription factor binding sites derived from transposable element sequences using ChIP-seq.* Methods Mol Biol, 2010. **674**: p. 225-40.

95. Wang, J., et al., *A c-Myc regulatory subnetwork from human transposable element sequences.* Mol Biosyst, 2009. **5**(12): p. 1831-9.

96. Polavarapu, N., et al., *Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA.* BMC Genomics, 2008. **9**: p. 226.

97. Marino-Ramirez, L., et al., *Transposable elements donate lineage-specific regulatory sequences to host genomes.* Cytogenet Genome Res, 2005. **110**(1-4): p. 333-41.

98. Wang, L., E.T. Norris, and I.K. Jordan, *Human Retrotransposon Insertion Polymorphisms Are Associated with Health and Disease via Gene Regulatory Phenotypes.* Front Microbiol, 2017. **8**: p. 1418.

99. Wang, L., et al., *Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements.* Nucleic Acids Res, 2017. **45**(5): p. 2318-2328.

100. Jjingo, D., et al., *Mammalian-wide interspersed repeat (MIR)-derived enhancers and the regulation of human gene expression.* Mob DNA, 2014. **5**: p. 14.

101. Huda, A., et al., *Prediction of transposable element derived enhancers using chromatin modification profiles.* PLoS One, 2011. **6**(11): p. e27513.

102. Marino-Ramirez, L. and I.K. Jordan, *Transposable element derived DNaseI-hypersensitive sites in the human genome.* Biol Direct, 2006. **1**: p. 20.

103. Wang, J., et al., *MIR retrotransposon sequences provide insulators to the human genome.* Proc Natl Acad Sci U S A, 2015. **112**(32): p. E4428-37.

104. Piriyapongsa, J., L. Marino-Ramirez, and I.K. Jordan, *Origin and evolution of human microRNAs from transposable elements.* Genetics, 2007. **176**(2): p. 1323-37.

105. Piriyapongsa, J. and I.K. Jordan, *A family of human microRNA genes from miniature inverted-repeat transposable elements.* PLoS One, 2007. **2**(2): p. e203.

106. Conley, A.B., W.J. Miller, and I.K. Jordan, *Human cis natural antisense transcripts initiated by transposable elements.* Trends Genet, 2008. **24**(2): p. 53-6.

107. Huda, A., et al., *Epigenetic regulation of transposable element derived human gene promoters.* Gene, 2011. **475**(1): p. 39-48.

108. Huda, A., et al., *Repetitive DNA elements, nucleosome binding and human gene expression.* Gene, 2009. **436**(1-2): p. 12-22.

109. Conley, A.B., J. Piriyapongsa, and I.K. Jordan, *Retroviral promoters in the human genome.* Bioinformatics, 2008. **24**(14): p. 1563-7.

110. Conley, A.B. and I.K. Jordan, *Cell type-specific termination of transcription by transposable element sequences.* Mob DNA, 2012. **3**(1): p. 15.

111. Cowley, M. and R.J. Oakey, *Transposable elements re-wire and fine-tune the transcriptome.* PLoS genetics, 2013. **9**(1): p. e1003234.

112. Kelley, D.R., et al., *Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions.* Genome biology, 2014. **15**(12): p. 537.

113. Shen, S., et al., *Widespread establishment and regulatory impact of Alu exons in human genes.* Proc Natl Acad Sci U S A, 2011. **108**(7): p. 2837-42.

114. Lev-Maor, G., et al., *The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons.* Science, 2003. **300**(5623): p. 1288-91.

115. Sorek, R., G. Ast, and D. Graur, *Alu-containing exons are alternatively spliced.* Genome research, 2002. **12**(7): p. 1060-1067.

116. Mersch, B., et al., *SERpredict: Detection of tissue-or tumor-specific isoforms generated through exonization of transposable elements.* BMC genetics, 2007. **8**(1): p. 78.

117. Anwar, S., W. Wulaningsih, and U. Lehmann, *Transposable elements in human cancer: causes and consequences of deregulation.* International journal of molecular sciences, 2017. **18**(5): p. 974.

118. Burns, K.H., *Transposable elements in cancer.* Nature Reviews Cancer, 2017. **17**(7): p. 415.

119. Clayton, E.A., et al., *Patterns of Transposable Element Expression and Insertion in Cancer.* Frontiers in Molecular Biosciences, 2016. **3**(76).

120. El Marabti, E. and I. Younis, *The Cancer Spliceome: Reprograming of Alternative Splicing in Cancer.* Frontiers in molecular biosciences, 2018. **5**: p. 80-80.

121. Jayasinghe, R.G., et al., *Systematic analysis of splice-site-creating mutations in cancer.* Cell reports, 2018. **23**(1): p. 270-281. e3.

122. Kahles, A., et al., *Comprehensive analysis of alternative splicing across tumors from 8,705 patients.* Cancer cell, 2018. **34**(2): p. 211-224. e6.

123. Oltean, S. and D.O. Bates, *Hallmarks of alternative splicing in cancer.* Oncogene, 2013. **33**: p. 5311.

124. Venables, J.P., *Aberrant and alternative splicing in cancer.* Cancer research, 2004. **64**(21): p. 7647-7654.

125. Venables, J.P., et al., *Cancer-associated regulation of alternative splicing.* Nature structural & molecular biology, 2009. **16**(6): p. 670.

126. Vitting-Seerup, K. and A. Sandelin, *The landscape of isoform switches in human cancers.* Molecular Cancer Research, 2017. **15**(9): p. 1206-1220.

127. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.

128. Tyner, C., et al., *The UCSC Genome Browser database: 2017 update.* Nucleic Acids Res, 2017. **45**(D1): p. D626-D634.

129. Smit, A., R. Hubley, and P. Green, *RepeatMasker Open-4.0. 2013–2015*. 2015.

130. Sondka, Z., et al., *The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers.* Nat Rev Cancer, 2018. **18**(11): p. 696-705.

131. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome biology, 2014. **15**(12): p. 550.

132. Shang, Z., et al., *Human kallikrein 2 (KLK2) promotes prostate cancer cell growth via function as a modulator to promote the ARA70-enhanced androgen receptor transactivation.* Tumor Biology, 2014. **35**(3): p. 1881-1890.

133. Adamopoulos, P.G., C.K. Kontos, and A. Scorilas, *Discovery of novel transcripts of the human tissue kallikrein (KLK1) and kallikrein-related peptidase 2 (KLK2) in human cancer cells, exploiting Next-Generation Sequencing technology*. Genomics, 2018.

134. David, A., et al., *Unusual alternative splicing within the human kallikrein genes KLK2 and KLK3 gives rise to novel prostate-specific proteins.* Journal of Biological Chemistry, 2002. **277**(20): p. 18084-18090.

135. Ma, Q., et al., *Identification and validation of key genes associated with non-small-cell lung cancer.* Journal of cellular physiology, 2019.

136. Sebestyén, E., M. Zawisza, and E. Eyras, *Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer.* Nucleic acids research, 2015. **43**(3): p. 1345-1356.

137. Castilla, L.H., et al., *The fusion gene CBFB-MYH11 blocks myeloid differentiation and predisposes mice to acute myelomonocytic leukaemia.* Nature genetics, 1999. **23**(2): p. 144.

138. Liu, P.P., et al., *Identification of the chimeric protein product of the CBFB-MYH11 fusion gene in inv (16) leukemia cells.* Genes, Chromosomes and Cancer, 1996. **16**(2): p. 77-87.

139. Castilla, L.H., et al., *Failure of Embryonic Hematopoiesis andLethal Hemorrhages in Mouse Embryos Heterozygousfor a Knocked-In Leukemia Gene CBFB–MYH11*. Cell, 1996. **87**(4): p. 687-696.

140. Ezponda, T., et al., *The histone methyltransferase MMSET/WHSC1 activates TWIST1 to promote an epithelial-mesenchymal transition and invasive properties of prostate cancer.* Oncogene, 2013. **32**(23): p. 2882-90.

141. Hudlebusch, H.R., et al., *The histone methyltransferase and putative oncoprotein MMSET is overexpressed in a large variety of human tumors.* Clinical Cancer Research, 2011. **17**(9): p. 2919-2933.

142. Keats, J.J., et al., *Overexpression of transcripts originating from the MMSET locus characterizes all t (4; 14)(p16; q32)-positive multiple myeloma patients.* Blood, 2005. **105**(10): p. 4060-4069.

143. Gerhardt, J., et al., *The androgen-regulated Calcium-Activated Nucleotidase 1 (CANT1) is commonly overexpressed in prostate cancer and is tumor-biologically relevant in vitro.* The American journal of pathology, 2011. **178**(4): p. 1847-1860.

144. Raney, B.J., et al., *Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser.* Bioinformatics, 2014. **30**(7): p. 1003-5.

145. Wunderlich, V., *Early references to the mutational origin of cancer.* Int J Epidemiol, 2007. **36**(1): p. 246-7.

146. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis.* Cell, 1990. **61**(5): p. 759-67.

147. Vogelstein, B. and K.W. Kinzler, *Cancer genes and the pathways they control.* Nat Med, 2004. **10**(8): p. 789-99.

148. Knudson, A.G., *Two genetic hits (more or less) to cancer.* Nat Rev Cancer, 2001. **1**(2): p. 157-62.

149. Boveri, T., *Zur Frage der Entstehung maligner Tumoren (The Origin of Malignant Tumors)(Jena: Gustav Fischer).* 1914.

150. Zhang, Y., et al., *A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases.* Cell Rep, 2018. **24**(2): p. 515-527.

151. Lee, T.I. and R.A. Young, *Transcriptional regulation and its misregulation in disease.* Cell, 2013. **152**(6): p. 1237-51.

152. Michalak, E.M., et al., *The roles of DNA, RNA and histone methylation in ageing and cancer.* Nature Reviews Molecular Cell Biology, 2019.

153. Ongen, H., et al., *Putative cis-regulatory drivers in colorectal cancer.* Nature, 2014. **512**(7512): p. 87-90.

154. Mayba, O., et al., *MBASED: allele-specific expression detection in cancer tissues and cell lines.* Genome Biol, 2014. **15**(8): p. 405.

155. Knight, J.C., *Allele-specific gene expression uncovered.* Trends Genet, 2004. **20**(3): p. 113-6.

156. Buckland, P.R., *Allele-specific gene expression differences in humans.* Hum Mol Genet, 2004. **13 Spec No 2**: p. R255-60.

157. Sigurdsson, M.I., et al., *Allele-specific expression in the human heart and its application to postoperative atrial fibrillation and myocardial ischemia.* Genome Med, 2016. **8**(1): p. 127.

158. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.

159. Liu, Z., X. Dong, and Y. Li, *A Genome-Wide Study of Allele-Specific Expression in Colorectal Cancer.* Front Genet, 2018. **9**: p. 570.

160. Sherr, C.J., *Principles of tumor suppression.* Cell, 2004. **116**(2): p. 235-46.

161. Sager, R., *Tumor suppressor genes: the puzzle and the promise.* Science, 1989. **246**(4936): p. 1406-12.

162. Flanagan, S.E., A.M. Patch, and S. Ellard, *Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations.* Genet Test Mol Biomarkers, 2010. **14**(4): p. 533-7.

163. the International Cancer Genome Consortium Mutation, P., et al., *Computational approaches to identify functional genetic variants in cancer genomes.* Nature Methods, 2013. **10**: p. 723.

164. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic acids research, 2003. **31**(13): p. 3812-3814.

165. Adzhubei, I., D.M. Jordan, and S.R. Sunyaev, *Predicting functional effect of human missense mutations using PolyPhen-2.* Current protocols in human genetics, 2013. **76**(1): p. 7.20. 1-7.20. 41.

166. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.

167. Knudson, A.G., *Mutation and cancer: statistical study of retinoblastoma.* Proceedings of the National Academy of Sciences, 1971. **68**(4): p. 820-823.

168. Khatami, F. and S.M. Tavangar, *A Review of Driver Genetic Alterations in Thyroid Cancers.* Iran J Pathol, 2018. **13**(2): p. 125-135.

169. Liu, Z., et al., *cisASE: a likelihood-based method for detecting putative cis-regulated allele-specific expression in RNA sequencing data.* Bioinformatics, 2016. **32**(21): p. 3291-3297.

170. Marsh, S.G., P. Parham, and L.D. Barber, *The HLA factsbook.* 1999: Elsevier.

171. Buhler, S. and A. Sanchez-Mazas, *HLA DNA sequence variation among human populations: molecular signatures of demographic and selective events.* PloS one, 2011. **6**(2): p. e14643-e14643.

172. Brandt, D.Y., et al., *Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data.* G3 (Bethesda), 2015. **5**(5): p. 931-41.

173. Gao, C., et al., *Identifying breast cancer risk loci by global differential allele-specific expression (DASE) analysis in mammary epithelial transcriptome.* BMC Genomics, 2012. **13**(1): p. 570.

174. Daelemans, C., et al., *High-throughput analysis of candidate imprinted genes and allele-specific gene expression in the human term placenta.* BMC genetics, 2010. **11**(1): p. 25.

175. Tuch, B.B., et al., *Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations.* PLoS One, 2010. **5**(2): p. e9317.

176. Hasin-Brumshtein, Y., et al., *Allele-specific expression and eQTL analysis in mouse adipose tissue.* BMC Genomics, 2014. **15**: p. 471.

177. Pastinen, T. and T.J. Hudson, *Cis-acting regulatory variation in the human genome.* Science, 2004. **306**(5696): p. 647-50.

178. Aguet, F., et al., *Local genetic effects on gene expression across 44 human tissues.* BiorXiv, 2016: p. 074450.

179. Albert, F.W. and L. Kruglyak, *The role of regulatory variation in complex traits and disease.* Nat Rev Genet, 2015. **16**(4): p. 197-212.

180. Duong, D., et al., *Using genomic annotations increases statistical power to detect eGenes.* Bioinformatics, 2016. **32**(12): p. i156-i163.

181. Consortium, G.T., *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.

182. Wagner, J.R., et al., *Computational analysis of whole-genome differential allelic expression data in human.* PLoS Comput Biol, 2010. **6**(7): p. e1000849.

183. Deaton, A.M. and A. Bird, *CpG islands and the regulation of transcription.* Genes Dev, 2011. **25**(10): p. 1010-22.

184. Costello, J.F., et al., *Aberrant CpG-island methylation has non-random and tumour-type–specific patterns.* Nature genetics, 2000. **24**(2): p. 132.

185. Bird, A.P., *CpG-rich islands and the function of DNA methylation.* Nature, 1986. **321**(6067): p. 209-213.

186. Siegfried, Z., et al., *DNA methylation represses transcription in vivo.* Nature genetics, 1999. **22**(2): p. 203.

187. Bird, A.P. and A.P. Wolffe, *Methylation-induced repression—belts, braces, and chromatin.* Cell, 1999. **99**(5): p. 451-454.

188. Kass, S.U., D. Pruss, and A.P. Wolffe, *How does DNA methylation repress transcription?* Trends in Genetics, 1997. **13**(11): p. 444-449.

189. Romanel, A., *Allele-Specific Expression Analysis in Cancer Using Next-Generation Sequencing Data.* Methods Mol Biol, 2019. **1878**: p. 125-137.

190. McClorey, G., et al., *Induced dystrophin exon skipping in human muscle explants.* Neuromuscul Disord, 2006. **16**(9-10): p. 583-90.

191. Lee, H.D., et al., *Exosome release of ADAM15 and the functional implications of human macrophage-derived ADAM15 exosomes.* FASEB J, 2012. **26**(7): p. 3084-95.

192. Kleino, I., R.M. Ortiz, and A.P. Huovila, *ADAM15 gene structure and differential alternative exon use in human tissues.* BMC Mol Biol, 2007. **8**: p. 90.

193. Bolger, J.C. and L.S. Young, *ADAM22 as a prognostic and therapeutic drug target in the treatment of endocrine-resistant breast cancer.* Vitam Horm, 2013. **93**: p. 307-21.

194. Ortiz, R.M., I. Karkkainen, and A.P. Huovila, *Aberrant alternative exon use and increased copy number of human metalloprotease-disintegrin ADAM15 gene in breast cancer cells.* Genes Chromosomes Cancer, 2004. **41**(4): p. 366-78.

195. Eswaran, J., et al., *RNA sequencing of cancer reveals novel splicing alterations.* Scientific Reports, 2013. **3**: p. 1689.

196. Salvador, F., et al., *Lysyl Oxidase–like Protein LOXL2 Promotes Lung Metastasis of Breast Cancer.* Cancer Research, 2017. **77**(21): p. 5846-5859.

197. Wu, L. and Y. Zhu, *The function and mechanisms of action of LOXL2 in cancer (Review).* Int J Mol Med, 2015. **36**(5): p. 1200-4.

198. da Silva, M.R., et al., *Splicing Regulators and Their Roles in Cancer Biology and Therapy.* BioMed research international, 2015. **2015**: p. 150514-150514.

199. Tsunoda, T., et al., *Involvement of Large Tenascin-C Splice Variants in Breast Cancer Progression.* The American Journal of Pathology, 2003. **162**(6): p. 1857-1867.

200. Guttery, D.S., et al., *Expression of tenascin-C and its isoforms in the breast.* Cancer and Metastasis Reviews, 2010. **29**(4): p. 595-606.

201. Hancox, R.A., et al., *Tumour-associated tenascin-C isoforms promote breast cancer cell invasion and growth by matrix metalloproteinase-dependent and independent mechanisms.* Breast cancer research : BCR, 2009. **11**(2): p. R24-R24.

202. Olivier, M., M. Hollstein, and P. Hainaut, *TP53 mutations in human cancers: origins, consequences, and clinical use.* Cold Spring Harbor perspectives in biology, 2010. **2**(1): p. a001008-a001008.

203. Lesueur, F., et al., *Single-nucleotide polymorphisms in the RB1 gene and association with breast cancer in the British population.* British journal of cancer, 2006. **94**(12): p. 1921-1926.

204. Kadam-Pai, P., et al., *Ethnic variations of a retinoblastoma susceptibility gene (RB1) polymorphism in eight Asian populations.* J Genet, 2003. **82**(1-2): p. 33-7.

205. Powell, S.M., et al., *APC mutations occur early during colorectal tumorigenesis.* Nature, 1992. **359**(6392): p. 235-237.

206. Knudson Jr, A.G. and L.C. Strong, *Mutation and cancer: neuroblastoma and pheochromocytoma.* American journal of human genetics, 1972. **24**(5): p. 514.

207. Knudson Jr, A.G. and L.C. Strons, *Mutation and cancer: a model for Wilms' tumor of the kidney.* Journal of the National Cancer Institute, 1972. **48**(2): p. 313-324.

208. Berger, A.H., A.G. Knudson, and P.P. Pandolfi, *A continuum model for tumour suppression.* Nature, 2011. **476**: p. 163.

209. Paige, A.J.W., *Redefining tumour suppressor genes: exceptions to the two-hit hypothesis.* Cellular and Molecular Life Sciences CMLS, 2003. **60**(10): p. 2147-2163.

210. Tucker, T. and J.M. Friedman, *Pathogenesis of hereditary tumors: beyond the "two-hit" hypothesis.* Clin Genet, 2002. **62**(5): p. 345-57.

211. Cheng, C.-W., et al., *Mechanisms of inactivation of E-cadherin in breast carcinoma: modification of the two-hit hypothesis of tumor suppressor gene.* Oncogene, 2001. **20**(29): p. 3814-3823.

212. Tomar, S., et al., *Mutation spectrum of RB1 mutations in retinoblastoma cases from Singapore with implications for genetic management and counselling.* PloS one, 2017. **12**(6): p. e0178776-e0178776.

213. Dommering, C.J., et al., *RB1 mutation spectrum in a comprehensive nationwide cohort of retinoblastoma patients.* Journal of medical genetics, 2014. **51**(6): p. 366-374.

214. Rushlow, D.E., et al., *Characterisation of retinoblastomas without RB1 mutations: genomic, gene expression, and clinical studies.* The lancet oncology, 2013. **14**(4): p. 327-334.

215. Slattery, M.L., et al., *The co-regulatory networks of tumor suppressor genes, oncogenes, and miRNAs in colorectal cancer.* Genes, chromosomes & cancer, 2017. **56**(11): p. 769-787.

216. Sung, J., et al., *Oncogene regulation of tumor suppressor genes in tumorigenesis.* Carcinogenesis, 2005. **26**(2): p. 487-494.

217. Goodarzi, H., O. Elemento, and S. Tavazoie, *Revealing global regulatory perturbations across human cancers.* Mol Cell, 2009. **36**(5): p. 900-11.

218. Cordero, D., et al., *Large differences in global transcriptional regulatory programs of normal and tumor colon cells.* BMC Cancer, 2014. **14**: p. 708.

219. Hill, C. and J. McDonald, *Evidence and potential clinical significance of changes in gene network interactions in ovarian cancer.* Journal of Biomedical Engineering and Informatics, 2015. **2**.

220. Jones, P.A. and S.B. Baylin, *The epigenomics of cancer.* Cell, 2007. **128**(4): p. 683-92.

221. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nat Genet, 2003. **33 Suppl**: p. 245-54.

222. Bickel, R.D., A. Kopp, and S.V. Nuzhdin, *Composite effects of polymorphisms near multiple regulatory elements create a major-effect QTL.* PLoS Genet, 2011. **7**(1): p. e1001275.

223. Mittal, V.K. and J.F. McDonald, *De novo assembly and characterization of breast cancer transcriptomes identifies large numbers of novel fusion-gene transcripts of potential functional significance.* BMC Med Genomics, 2017. **10**(1): p. 53.

224. Havens, M.A. and M.L. Hastings, *Splice-switching antisense oligonucleotides as therapeutic drugs.* Nucleic acids research, 2016. **44**(14): p. 6549-6563.

225. Liemberger, B., et al., *RNA Trans-Splicing Modulation via Antisense Molecule Interference.* Int J Mol Sci, 2018. **19**(3).

226. Quinlan, A.R., *BEDTools: the Swiss-army tool for genome feature analysis.* Current protocols in bioinformatics, 2014. **47**(1): p. 11.12. 1-11.12. 34.

227. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016. **17**(1): p. 122.

228. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.

229. The Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**: p. 68.

230. Koboldt, D.C., et al., *VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing.* Genome Res, 2012. **22**(3): p. 568-76.

231. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2011. **39**(Database issue): p. D52-7.

232. Hodgkinson, A., et al., *A haplotype-based normalization technique for the analysis and detection of allele specific expression.* BMC Bioinformatics, 2016. **17**(1): p. 364.

233. Degner, J.F., et al., *Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data.* Bioinformatics, 2009. **25**(24): p. 3207-12.

234. Derrien, T., et al., *Fast computation and applications of genome mappability.* PloS one, 2012. **7**(1): p. e30377.

235. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.

236. Delaneau, O., J. Marchini, and J.F. Zagury, *A linear complexity phasing method for thousands of genomes.* Nat Methods, 2011. **9**(2): p. 179-81.

237. Shen, R. and V. Seshan, *FACETS: Allele-Specific Copy Number and Clonal Heterogeneity Analysis Tool Estimates for High-Throughput DNA Sequencing.* 2016.

238. Bansal, V. and V. Bafna, *HapCUT: an efficient and accurate algorithm for the haplotype assembly problem.* Bioinformatics, 2008. **24**(16): p. i153-9.

239. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* The American Journal of Human Genetics, 2007. **81**(3): p. 559-575.

240. Gadin, J.R., et al., *AllelicImbalance: an R/bioconductor package for detecting, managing, and visualizing allele expression imbalance data from RNA sequencing.* BMC Bioinformatics, 2015. **16**: p. 194.

241. You, B.H., S.H. Yoon, and J.W. Nam, *High-confidence coding and noncoding transcriptome maps.* Genome Res, 2017. **27**(6): p. 1050-1062.

242. Stephens, R.M. and T.D. Schneider, *Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites.* J Mol Biol, 1992. **228**(4): p. 1124-36.

243. Prasad, V., T. Fojo, and M. Brada, *Precision oncology: origins, optimism, and potential.* The Lancet Oncology, 2016. **17**(2): p. e81-e86.

244. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.

245. Ayers, M., et al., *Gene expression profiles predict complete pathologic response to neoadjuvant paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide chemotherapy in breast cancer.* Journal of clinical oncology, 2004. **22**(12): p. 2284-2293.

246. Collins, I. and P. Workman, *New approaches to molecular cancer therapeutics.* Nature Chemical Biology, 2006. **2**(12): p. 689.

247. Ross, J.S. and J.A. Fletcher, *The HER-2/neu Oncogene in Breast Cancer: Prognostic Factor, Predictive Factor, and Target for Therapy.* Stem cells, 1998. **16**(6): p. 413-428.

248. Costello, J.C., et al., *A community effort to assess and improve drug sensitivity prediction algorithms.* Nat Biotechnol, 2014. **32**(12): p. 1202-12.

249. Geeleher, P., N.J. Cox, and R.S. Huang, *Clinical drug response can be predicted using baseline gene expression levels and in vitro drug sensitivity in cell lines.* Genome Biol, 2014. **15**(3): p. R47.

250. Hejase, H. and C. Chan, *Improving drug sensitivity prediction using different types of data.* CPT: pharmacometrics & systems pharmacology, 2015. **4**(2): p. 98-105.

251. Suphavilai, C., D. Bertrand, and N. Nagarajan, *Predicting Cancer Drug Response Using a Recommender System.* bioRxiv, 2017: p. 215327.

252. Sekula, M.N., *OptCluster: an R package for determining the optimal clustering algorithm and optimal number of clusters.* 2015.

253. Liaw, A. and M. Wiener, *Classification and regression by randomForest.* R news, 2002. **2**(3): p. 18-22.

254. Mi, H., et al., *Large-scale gene function analysis with the PANTHER classification system.* Nature protocols, 2013. **8**(8): p. 1551.

255. Hoadley, K.A., et al., *Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer.* Cell, 2018. **173**(2): p. 291-304. e6.

256. Komiya, Y. and R. Habas, *Wnt signal transduction pathways.* Organogenesis, 2008. **4**(2): p. 68-75.

257. Chiurillo, M.A., *Role of the Wnt/β-catenin pathway in gastric cancer: An in-depth literature review.* World journal of experimental medicine, 2015. **5**(2): p. 84.

258. Moreno-Layseca, P., et al., *Integrin trafficking in cells and tissues.* Nature cell biology, 2019: p. 1.

259. Seguin, L., et al., *Integrins and cancer: regulators of cancer stemness, metastasis, and drug resistance.* Trends in cell biology, 2015. **25**(4): p. 234-240.

260. Hutson, M., *Artificial intelligence faces reproducibility crisis*. 2018, American Association for the Advancement of Science.

261. Lloyd, S., *Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM.* IEEE Trans. Inform. Theor.(1957/1982) Google Scholar, 1957.

262. Johnson, S.C., *Hierarchical clustering schemes.* Psychometrika, 1967. **32**(3): p. 241-254.

263. Rousseeuw, P.J. and L. Kaufman, *Finding groups in data.* SeriesinProbability&MathematicalStatistics㎡ 1990㎡34 (1), 1990: p. 111-112.

264. Banfield, J.D. and A.E. Raftery, *Model-based Gaussian and non-Gaussian clustering.* Biometrics, 1993: p. 803-821.

265. Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.* Proceedings of the National Academy of Sciences, 1999. **96**(6): p. 2907-2912.

266. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic acids research, 2014. **43**(D1): p. D805-D811.

267. Frankish, A., et al., *GENCODE reference annotation for the human and mouse genomes.* Nucleic Acids Res, 2019. **47**(D1): p. D766-D773.

268. Kahles, A., et al., *SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data.* Bioinformatics, 2016. **32**(12): p. 1840-1847.

269. Kent, W.J., et al., *The human genome browser at UCSC.* Genome research, 2002. **12**(6): p. 996-1006.